THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

The characterization of peptides expressed from short open reading frames (sORFs)

Ting-Yu, Feng (Caroline)

Master of Molecular Bioscience

A thesis submitted for the degree of Master of Philosophy at

The University of Queensland in 2014

School of Chemistry and Molecular Biosciences

Abstract

Potentially translatable short open reading frames (sORFs) of less than 100 codons are present on both mRNAs and non-coding RNAs (ncRNAs) and 50% of mammalian mRNAs contain at least one sORF. We hypothesize that a subset of sORFs encode for functional short peptides (sPEPs) that are expressed and contribute to proteome complexity. Our recent bioinformatic studies showed that nearly 2% of sORFs are conserved between several species indicating that these sORFs may have critical functions. Recent proteomic studies have identified over 1,000 sPEPs in human cell lines showing that some sORFs are indeed translated. Surprisingly, a number of peptides have been identified that are encoded by sORFs present in ncRNAs. In order to extend and validate these studies, I extracted low molecular weight proteins from HeLa and HEK293 cell lysates by either SDS-PAGE or ERLIC fractionation. These extracts were digested with trypsin or LysC and analysed by nano LC-MS/MS. The resulting MS/MS data was searched against the UniProKB/Swiss-Prot using MASCOT version 2.4 to filter out known proteins, and all unmatched spectra were searched against the human RefSeq database. ProteinPilot^{TM} was also used to identify sORF-encoded peptides by searching against an in-house sORF and the Human Alternative Open Reading Frame (HaltORF) databases. To date, I have identified several sPEPs including three that are novel. These sPEPs have a mass of less than 20 kDa. One of those is expressed from an ncRNA transcript and is expected to be secreted. The other two sPEPs are encoded by upstream open reading frames (uORFs) on mRNA transcripts; one of these is predicted to localise to the cytoplasm while the other is expected to be secreted. The role, if any, of these peptides has yet to be determined but their identification has provided a pool of candidates for further molecular characterization in order to determine their function.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

No publications

Publications included in this thesis

No publications included

Contributions by others to the thesis

No contributions by others

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

First of all, I would like to express my utmost gratitude to my supervisors, Dr Amanda Nouwens, Assoc Prof Joe Rothnagel, and Prof Ross Smith, for the great guidance and support throughout my research career. This thesis would not have been possible without their unlimited help and encouragement. A very special thanks to Dr Amanda Nouwens who has been dedicating her time in guiding me all the way from the peptide enrichment strategies to the operation of the Mass Spectrometry instrument. It is an honour for me to work with her assistant and suggestions for hurdling all the obstacles in my research work.

I am heartily thankful to my principal supervisor, Assoc Prof Joe Rothnagel, who has made available his support in a number of ways. Another big thanks goes out to Professor Ross Smith for his support and advices. I would also like to take this opportunity to record my sense of gratitude to all the members of the Smith & Rothnagel's lab for sharing their knowledge and providing such joyous environment to work in.

Lastly but not the least, I offer my regards and blessing to my family and friends who supported and inspired me in any respect.

Keywords

short open reading frames,non-coding RNAs, ERLIC, SDS-PAGE, LC-MS/MS

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060109, Proteomics and Intermolecular Interactions (excl. Medical Proteomics), 70%

ANZSRC code: 060102, Bioinformatics, 30%

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 100%

List of Figures & Tables

Figures

Tables

List of abbreviations

ABC	Ammonium bicarbonate
ACN	Acetonitrile
BLAST	Basic Local Alignment Search Tool
CDS	Coding sequence
DTT	Dithiothreitol
ESI-Q-TOF-MS	Electrospray Standard Ionization-Quadrapole-Time Of Flight Mass Spectrometry
GeLC-MS/MS	Gel electrophoresis separation-LC-MS/MS
HEK293	Human Embryonic Kidney 293 cells
HeLa	Cervical cancer cell line from Henrietta Lacks
IAA	Iodoacetamide
LC-MS/MS	Liquid chromatography coupled with tandem mass spectrometry
mCDS	Main coding sequence
uAUG	Upstream AUG

SDS-PAGE	Sodium dodecyl sulfate- polyacrylamide gel electrophoresis
sORF	Collective term for all small open reading frames that are generally less than 100 codons
sPEP	Peptide encoded by small open reading frames
oORF	Overlappingopen reading frame
oPEP	Small peptide that overlaps the mCDS
uORF	Upstream open reading frame
dPEP	Peptides encoded by downstream open reading frames
ncRNA	Non-coding RNA
ncPEP	Peptides encoded by sORFs present on non-coding RNAs
uPEP	Upstream peptide
MS/MS	Tandem mass Liquid chromatography coupled with tandem mass spectrometry
MWCO	Molecular Wight Cut-Off
NCBI	National Center for Biotechnology Information
ORF	Open reading frame
RefSeq	Reference Sequence database

Table of Contents

Chapter 1

Introduction

Translation initiation of eukaryotic mRNAs is via the scanning mechanism, which starts with the migration of the pre-initiation complex along the 5'- untranslated region (5'UTR) until an appropriate AUG codon is reached (Wang and Rothnagel, 2004). Specific eukaryotic initiation factors (eIFs) including eIF2-GTP/Met- tRNA, eIF1A, and eIF3, and the initiator methionine-tRNA carried by the 40S ribosome subunits form a pre-initiation complex, known as the 43S complex (Rogozin et al., 2001). The 43S complex scans the mRNA along the 5' UTR in a 5' to 3' direction until an AUG start codon is encountered. At this stage, 60S subunits are involved the formation of an 80S ribosome that allows the decoding of RNA into protein (Rogozin et al., 2001) (Figure 1.1). Translation of most eukaryotic mRNAs are initiated at the first AUG triplet in the 5' UTR (Rogozin et al., 2001). However, in some cases, alternative initiation of translation at different start codons exists, especially for some growth factor genes and proto-oncogenes (Willis, 1999).The most efficient context for translational initiation at a AUG start codon is known as the Kozak sequence (GCCA/**G**CCAUG**G**), which is a consensus sequence for translational initiation of eukaryotic mRNAs (Crowe et al., 2006). The most critical positions within this sequence are positions -3(A or G) and +4 (G) which determine the strength of the initiator and translational efficiency (Crowe et al., 2006).

There are two mechanisms involved in the ribosomal scanning models that modify the translation of the main ORF down-stream with one or more uAUGs or uORFs (Wang and Rothnagel, 2004): 1) The leaky scanning mechanism (Figure 1.2c). 2) and the re-initiation mechanism (Figure 1.2b). The leaky scanning mechanism occurs when some 40S subunits do not recognize every uAUG codon, as a result, some ribosomes will then initiate at the downstream AUG codon (Wang and Rothnagel, 2004). The re-initiation mechanism is thought to be inefficient since it only occurs in short uORFs after their translation (Wang and Rothnagel, 2004), where the 40S subunit remains bound to the mRNA after translation and continues to initiate at a downstream AUG codon (Meijer and Thomas,

2002). The relatively short time between the initiation and termination is thought to be able to induce re-binding between initiation factors and the 40S ribosomal subunit. As the distance between two ORFs increases, the time for reloading the 40S subunit will then increase, resulting in the enhancement of re-initiation efficiency (Meijer and Thomas, 2002).

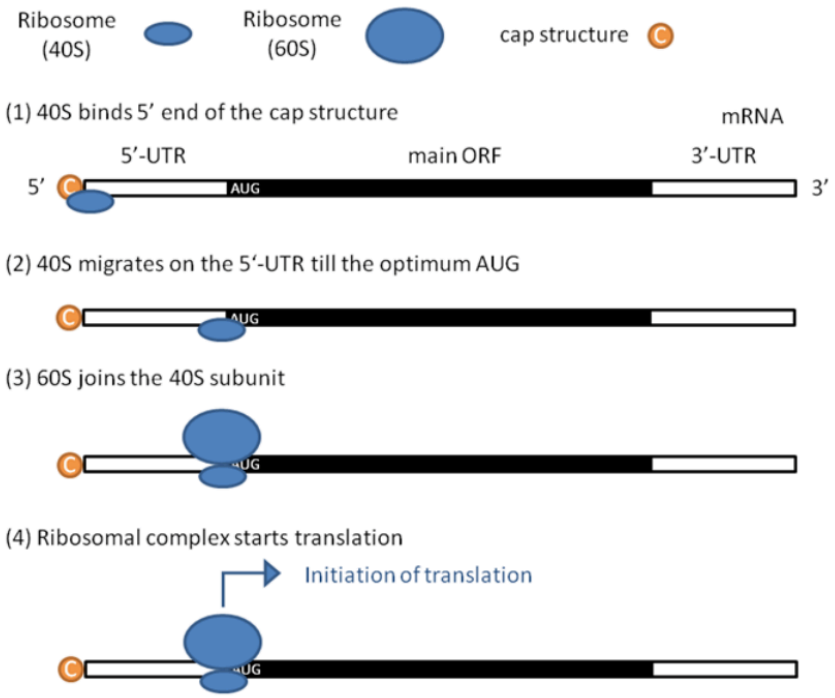


Figure 1.1. The mechanism of translation initiation on eukayrotic mRNAs.

The attachment of 40S ribosomal subunit at the 5'-end of the mRNA and then migrates along the 5'-UTR until an AUG initiator codon is encountered. A large 60S subunit binds to the 40S subunit and the complete ribosomal complex triggers translation for protein synthesis (Ao-Kondo et al., 2011).

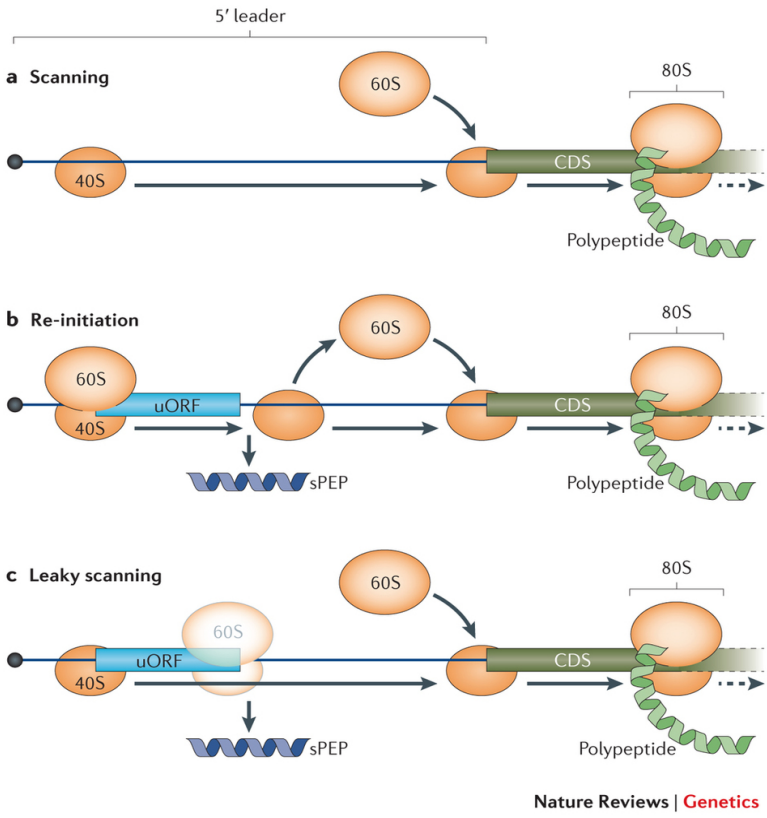


Figure 1.2. The irregular model of ribosomal scanning mechanisms on eukaryotic mRNAs.

a) The regular model of ribosomal scanning mechanism on eukaryotic mRNAs. The 40S complex scans the mRNA along the 5' UTR in a 5' to 3' direction until an AUG start codon is encountered. At this stage, 60S subunits are involved the formation of an 80S ribosome that allows the decoding of RNA into protein (Rogozin et al., 2001). b) Re-initiation mechanism is initiated when the 40S subunit remains to the mRNA after translation and continues to initiate at a downstream AUG codon. The relatively short time between the initiation and termination is thought to be able to induce re-binding between initiation factors and the 40S ribosomal subunit (Meijer and Thomas, 2002). c) Leaky scanning occurs when some 40S subunits do not recognize every uAUG codon, as a result, some ribosomes will then initiate at the start AUG (sAUG) codon (Wang and Rothnagel, 2004).

Short open reading frames (sORFs) are pervasive in eukaryote genomes but only a subset are likely to be translated. Recent innovations in computing, proteomics and high throughput analysis of translation start sites have sparked a renewed interest in open reading frames between 10 and 100 codons in size.

Characterisation of sORFs by location

On average, sORFs constitute about 5% of all annotated ORFs in the NCBI RefSeq database for a variety of eukaryotes including mammals (Kastenmayer et al., 2006). Short ORFs can occur by chance throughout the genome yet one study found that most sORFs (94%) are present in regions that are transcribed (Frith et al., 2006), indicating a high potential for expression of their encoded peptides. The common identifying characteristic of sORFs is the length of their open reading frames. Theoretically, these could be as small as 3 codons but the smallest translated sORF described to date is 6 codons (Law et al., 2001). The upper limit of sORFs has been arbitrarily set at 100 codons largely as a consequence of gene prediction algorithms ignoring open reading frames smaller than this (Basrai et al., 1997, Claverie, 1997). Clearly there are likely to be sORFs that extend past this artificial limit and indeed a recent proteomic study has identified peptides encoded by sORFs of up to 250 codons in length (Slavoff et al., 2013). There are five subcategories of sORFs (Figure 1.3): sORFs that are located on a variety of non-coding RNAs (ncRNAs: long, intergenic and anti-sense); upstream open reading frames (uORFs) located within the 5′ UTR of mRNAs; downstream open reading frames (dORFs) located within the 3′ UTR; short overlapping ORFs (oORFs) that are in the mCDS or located out of the mCDS in non-canonical +2 and +3 open reading frames.

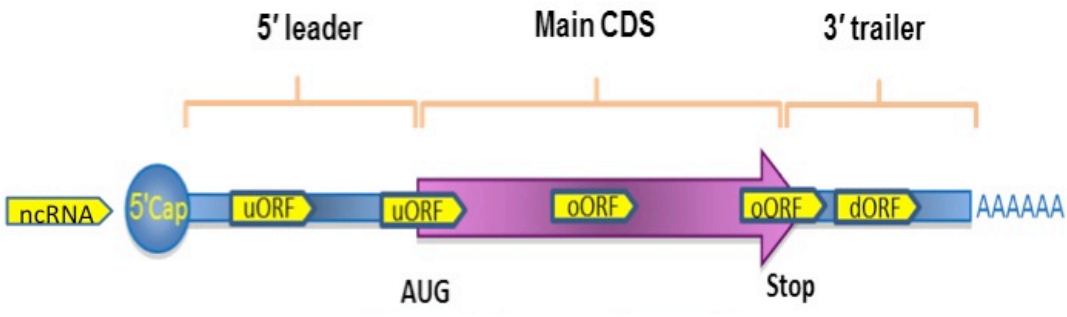


Figure 1.3. Schematic showing the various geographic locations of sORFs.

There are five subcategories of sORFs (shown in boxes): sORFs that are located on a variety of non-coding RNAs (ncRNAs: long, intergenic and anti-sense); upstream open reading frames (uORFs) located within the 5′ UTR (5′ leader) of mRNAs; downstream open reading frames (dORFs) located within the 3′ UTR (3′ trailer); short overlapping ORFs (oORFs) that are in the mCDS or t located out of the mCDS in non-canonical +2 and +3 open reading frames.

Small coding sequence in 5'UTRs (uORFs)

uORFs refer to one or more ORFs located in the 5'UTR prior to the main open reading frame (mORF). These uORFs are regarded as important elements involved in transcriptional regulation in the *cis*-regulation of gene expression (Wethmar et al., 2010). The influence of individual uORFs on translation of the downstream mORF is determined by characteristics such as length, number per transcript, secondary context and distance to the mORF (Wethmar et al., 2010).

The occurrence of at least one uAUG contained in vertebrate mRNAs has been shown to range from 11% to 42% depending on the species (Kozak, 1987, Pesole et al., 2000). For human specifically, 20%-49% of mRNAs have been found to contain at least one uAUG (Pesole et al., 2000, Davuluri et al., 2000, Barbosa et al., 2013). Several studies showed that the presence of uAUGs/uORFs would reduce the number of ribosomes that initiates the AUG start codons and subsequently diminishes the efficiency of translational initiation on the mORF (Meijer and Thomas, 2002,

Davuluri et al., 2000). The conservation of uORFs has been reported in several species such as human, mice, yeast, plants and insects (Table A1).

Small coding sequence in non-coding RNAs (ncRNAs)

sORFs also appear at high frequency within ncRNAs (Slavoff et al., 2013). It is believed that short ncRNAs (~50 nt) are too small to be translated so most translatable sORFs are generally thought to be found on long ncRNAs (lncRNAs) (~200 nt) (Kageyama et al., 2011). However, ribosomal profiling has been reported to detect translation initiation sites (TIS) on short ncRNAs (Lee et al., 2012). Around 600,000 potential ncRNAs had been identified in *Arabidopsis* with the analysis of intergenic (Hanada et al., 2007) and whole genome sequences (Lease and Walker, 2006). Identification of ncRNAs in different species, such as *Drosophila* (Ladoukakis et al., 2011), plant (Yang et al., 2011), mice (Frith et al., 2006) and human (Slavoff et al., 2013) have also been reported.

Small coding sequence in mRNAs: oORFs, dORFs

oORFs has been differentiated into a number of subtypes: those that sit within a mORF; those that extend from the mORF to 3'UTR and dual-coding transcripts produced by alternative splicing (Ribrioux et al., 2008, Michel et al., 2012). The conservation of oORFs between human and mice have been reported in several studies (Chung et al., 2007, Ribrioux et al., 2008, Xu et al., 2010).

Unlike the 5′UTR and coding regions, the 3′UTR was considered not to be translated so less emphasis was made to identify and characterise downstream ORFs (dORFs) (Ingolia et al., 2011). Nevertheless, 3′UTRs are much longer than 5′UTRs so potentially could contain more ORFs (Mercer et al., 2011). Although less attention has been paid to dORFs because of a lack of apparent functionality, there is accumulating evidence that some may be translated (Ingolia et al., 2011).

As out-of frame alternative translation initiation have been found to encode proteins of different amino acid composition in viruses and bacteriophages (Normark et al., 1983), recently studies reported the discovery of several protein products encoded from oORFs that are located in UTRs or overlapping mCDSs in non-canonical +2 and +3 open reading frames (Vanderperre et al., 2013). A recent study detected 1,259 alternative proteins, indicating that oORFs are indeed translated and contribute to human proteome (Vanderperre et al., 2013).

Evidence of sORF translation from proteomic studies

Peptides encoded by uORFs: uPEPs

Proteomic studies on human cell isolates have reported that some uORFs are indeed translated to uPEPs (Oyama et al., 2004, Slavoff et al., 2013,Oyama et al., 2007). The first evidence for uPEPs using high-resolution nanoflow liquid chromatography equipped with electrospray ionization tandem mass spectrometry in human cells was published (Oyama et al., 2007)and they identified eight uPEPs in those studies. A similar proteomic approach was also performed using a combination of a more sensitive chromatographic method, electrostatic repulsion-hydrophilic interaction chromatography (ERLIC), to fractionate the peptide mixture together with RNA-seq transcriptome data (Slavoff et al., 2013). This resulted in the identification 90 proteins translated from sORFs, 22 of which were derived from the translation of uORFs (Slavoff et al., 2013).

Peptides encoded by sORFs present in ncRNAs

A proteomic study identified 5426 short peptides on *Arabidopsis*, 905 of which encoded by genes had not been previously annotated in the reference databases (Castellana et al., 2008). Slavoff *et al*. identified 49 sPEPs encoded from ncRNAs and 8 sPEPs from intergenic ncRNAs (lincRNAs) in human cell lines.

Peptides encoded by oORFs and dORFs: oPEPs and dPEPs

A recent paper has identified 11 oPEPs (Slavoff et al., 2013) in human cell lines. Another proteomic study detected 1,259 alternative proteins in human cell lines, tissues, and fluids, indicating that oORFs are indeed translated and contribute to human proteome (Vanderperre et al., 2013).

A few studies focusing on the identification of dPEPs have been published. A study reported that dORFs could be translated via leaky scanning and ribosomal reinitiation mechanisms or translated as a ribosomal entry site contained on the 3′UTR (Mercer et al., 2011). Three dPEPs translated from two dORFs were reported in a proteomic study (Oyama et al., 2007) and six more dPEPs were found in a recent proteomic paper (Slavoff et al., 2013).

Bioinformatic identification of sPEPs

In order to find functional sPEPs, searches for conservation between species at the amino acid level play an important role because cross-species conservation could indicate that the sequence is maintained evolutionally from a functional role. sORFs that lack cross-species conservation are unlikely to encode functional peptides (Andrews and Rothnagel, 2014). However, these sORFs can not be disregarded because they may be biologically relent to species-specific sPEPs. Bioinformatic studies have reported a large amount of sORFs with functional potential and which may need further investigation for confirmation (Table A2). Once sORFs is verified with its capability, it implies that the undiscovered sPEPs existing in the sORFs contribute to proteome complexity. In addition, a number of sPEPs have been identified from sORFs in ncRNAs and in lincRNAs (Slavoff et al., 2013), providing reliable evidence that human proteome is much more complicated than previously appreciated.

Ribosome profiling and identification of occupied start codons

Ribosome profiling strategies have recently emerged as a powerful tool to map which mRNA transcripts are translated at any particular stage and at what efficiency based on deep sequencing of ribosomal protected mRNA transcripts (Kuersten et al., 2013). Ribosomal profiling has been reported to detect translation initiation sites (TIS) on short ncRNAs (Lee et al., 2012). Support for sORF translation is provided by ribosomal profiling studies, which generate a transcriptome-wide map of translation initiation sites. A recent study on human cells identified 4400 translation initiation sites that matched uORFs (Vanderperre et al., 2012) many of which show conservation in other species. Ribosomal footprinting technique has been developed based upon high-throughput DNA sequencing that provides systematic monitoring of protein translation in mammalian cells (Ingolia et al., 2011). To validate human TISs from the previous published uORFs, 14 of the start sites identified by ribosomal footprinting had been classified as demarcating functional uORFs (Fritsch et al., 2012)

Identification of functional sPEPs

uPEPs

Despite that several uORFs are known to be translated, only a few have been identified to have functional roles (Table A3). For example, in *Arabidopsis thaliana,* expression of the CDS is regulated by polyamines binding to the nascent upstream sPEP; orthologous to human *SAMDC1* plasmid (Hanfrey et al., 2005). In human, a sPEP has been found to be implicated in the regulation of human hairles homolog (HR*)*; 13 causative mutations of Marie Unna hereditary hypotrichosis have been identified within the second uORF (Wen et al., 2009*)*. A subset of these attenuate translation of the downstream ORF in response to environmental signals, termed “peptoswitch” (Jorgensen and Dorantes-Acosta, 2012). The regulatory role of the peptoswitch is to activate uPEPs

to bind to small molecules through an intermediary (Jorgensen and Dorantes-Acosta, 2012). A recent research group reported that the expression of human microsomal epoxide hydrolase (EPHX1), a critical xenobiotic-metabolizing enzyme, catalyzing both detoxification and bioactivation reactions, was inhibited by *trans*-acting sPEPs that were encoded by two uORFs through interactions with the translation machinery (Nguyen et al., 2013*)*. The majority of uPEPs function in *cis*-regulational translation of downstream ORF through identified mechanisms (Wen et al., 2009, Hanfrey et al., 2005).

Peptides encoded from ncRNAs

Several sPEPs on intergenic regions and ncRNAs, particularly in plants and insects, have been identified to be functional (Table A3). Those sPEPs have been shown to have various regulatory roles, although the functional mechanisms underlying their roles are still under investigation. Two examples of functional mechanisms underlying the sPEPs encoded from intergenic and ncRNAs have been reported in *D. melanogaster* (Magny et al., 2013, Kondo et al., 2010). Peptides of 11-32 amino acids encoded by the polished rice (pri) on a long ncRNA has been found to control epidermal differentiation in *D. melanogaster* by triggering the amino-terminal truncation of the Shavenbaby (Svb) protein, thereby converting Svb from a repressor to an activator (Kondo et al., 2010). Pri, therefore, plays an important role in providing a strict control in epidermal morphogenesis (Kondo et al., 2010). As a result of this study, Pri sORF have been reannotated as a mRNA. Another research group described two peptides of less than 30 amino acids encoded by the *sarcolamban* locus of *Drosophila*. These peptides regulate the calcium transport by associating with sarco-endoplasmic reticulum Ca^{2+} adenosine triphosphatase (SERCA), and hence affecting in regular muscle contraction in *Drosophila* heart (Magny et al., 2013).

oPEPs/ dPEPs

Apart from uPEPs and ncRNAs, a number of oPEPs have been reported to be functional, particularly in plants and insects (Frank and Smith, 2002, Röhrig et al., 2002, Narita et al., 2004, Colombani et al., 2012). Only a few human sPEPs encoded from overlapping ORFs have been characterised so far (Table A3). An oPEP expressed from the intestinal carboxyl esterase gene has been reported to be recognized by human leukocyte antigen-B7-restricted renal cell carcinoma-reactive T cell clone by binding HLA-B*0702-presenting molecules (Ronsin et al., 1999). This peptide product may be associated with regulation of gene expression and cancer by binding to Ag-presenting molecules and involving in Ag-processing mechanism (Ronsin et al., 1999). Understanding the mechanisms of the translation of oORFs in tumor cells is important in the investigation of tumor immunology. Recent studies have characterised two oPEPs encoded from alternative proteins in human cell lines, AltPrP (Vanderperre et al., 2011) and AltATXN1 (Bergeron et al., 2013). Although only a few dPEPs have been identified, they are found to be translated from dORFs via leaky scanning and ribosomal reinitiation mechanisms or translated as a ribosomal entry site contained on the 3 ´ UTR (Mercer et al., 2011). A recent proteomic study reported a dPEP encoded from a sORF within the 3′trailer sequence of the murine retrovirus integration site 1 homologue (MRVI1) gene (Vanderperre et al., 2013). The AltMRVI1 sPEP was found to colocalise with the breast cancer type 1 susceptibility protein (BRCA1) in the nucleus, and the interaction between them was also confirmed through co-immunoprecipitation (Vanderperre et al., 2013). However, the role of AltMRVI1 associated in this interaction remains unkown.

Aims and Significance

Physical evidence for sORF-encoded peptides has come from proteomic studies on human cell isolates using 2D nano-liquid chromatography-tandem mass spectrometry (Oyama et al., 2004,

Oyama et al., 2007, Vanderperre et al., 2013, Slavoff et al., 2013). We hypothesized that sORFs encode functional peptides and are endogenously expressed as part of the eukaryotic cellular proteome, contributing to proteome complexity. We believe sPEPs have biological roles beyond their canonical function, such as having trans-acting roles in other gene regulation pathways or being involved in cell development and function.

Our recent bioinformatic studies showed that nearly 2% of sORFs are conserved between several species indicating that these sORFs may have critical functions. Recent proteomic studies have identified over 1,000 sPEPs in human cell lines showing that some sORFs are indeed translated. Surprisingly, a number of peptides have been identified that are encoded by sORFs present in ncRNAs. In order to extend and validate these studies, we extracted low molecular weight proteins from HeLa and HEK293 cell lysates to confirm our hypothesis that sPEPs do exist and have distinct biological roles in cells.

In this project, the experimental methodology was divided into three parts, which were illustrated in the flowchart in Figure 1.4.

Aim1

The first step of this project was to validate peptide enrichment methods to reduce sample complexity. Low molecular weight proteins from HeLa and HEK293 cell lysates were extracted by either SDS-PAGE or ERLIC fractionation. These extracts were digested with trypsin or LysC and analysed by nano LC-MS/MS.

Aim2

The second step was to research the resulting MS/MS data against the UniProKB/Swiss-Prot using MASCOT version 2.4 to filter out known proteins, and all unmatched spectra were searched against

the human RefSeq database. ProteinPilot™was also used to identify sORF-encoded peptides by searching against an in-house sORF and the HaltORF databases.

Aim3

Bioinformatic analyses for confirmed sPEPs from the proteomic approach were performed before characterisation. Searches for cross-species conservation of sORFs can reveal those that encode potential functionally important peptides. High levels of sequence identity between sORF homologues are an indication that the encoded uPEP has been maintained during evolution.

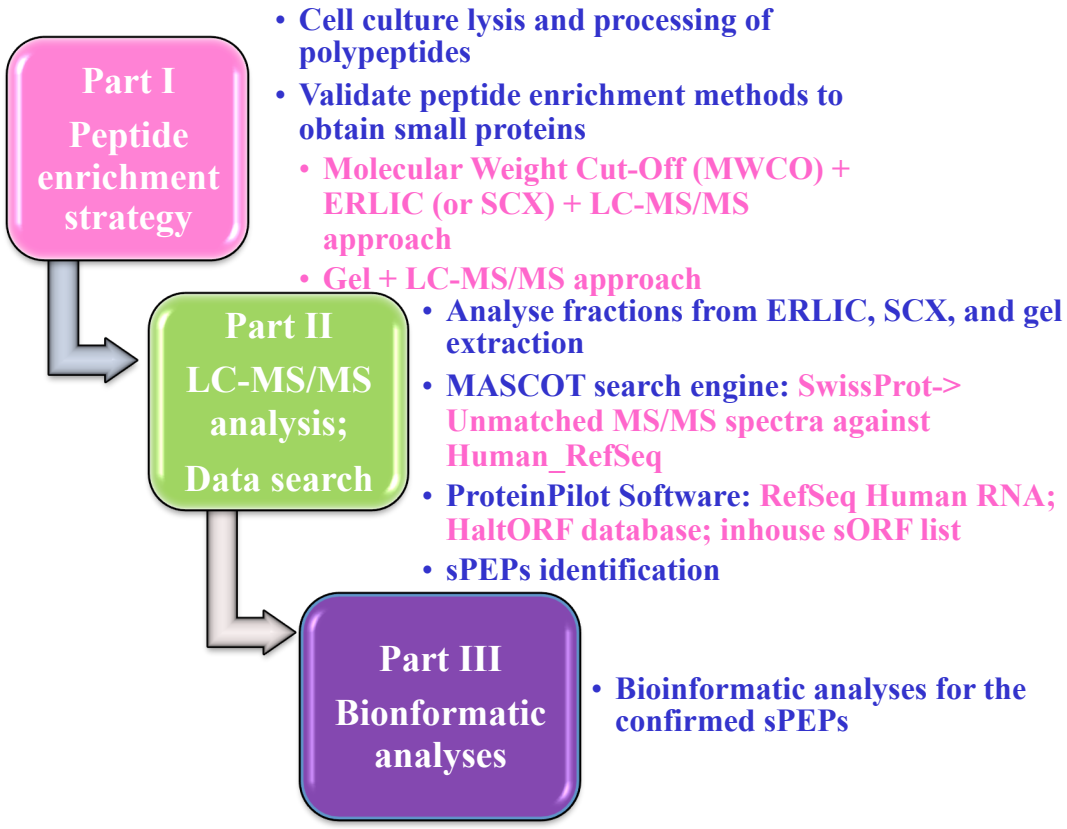


Figure 1.4. The overall experimental methodology for this research project.
The flowchart consists of three main parts. Part I involves protein enrichment strategies used to separate polypeptides. Part II includes the development of targeted MS analysis. Part III focuses on bioinformatic analyses for identified sPEPs.

Chapter 2

Materials and Methods

Materials

HEK293 cells / HeLa cells

DMEM supplemented with L-glutamine, 10% v/v newborn calf serum (NCS)

Penicillin

Streptomycin

Trypsin/EDTA solution

Phosphate buffered saline (PBS)

Bortezomib (PS-341) proteasome inhibitor (PI)

Dimethyl sulfoxide (DMSO)

0.25% Acetic acid

10 kDa, 30 kDa, and 50 kDa molecular weight cut-off (MWCO) filters (Amicon Ultra Kit)

2D Quant Kit

Dithiothreitol (DTT)

Iodoacetamide (IAA)

Trypsin/ LysC digestion solution

Acetonitrile (ACN)

Electrostatic repulsion hydrophilic interaction chromatography (ERLIC)

Strong cationic exchange (SCX)

SDS-PAGE gel electrophoresis

LC-MS/MS

MASCOT version 2.4/ UniProKB/Swiss-Prot search, Matrix Science Limited

ProteinPilot^{TM} (ABSCIEX)

Cell culture and lysate

HEK293 and HeLa cells were cultured in high-glucose DMEM supplemented with L-glutamine, 10% v/v newborn calf serum (NCS), 100 µg/ml penicillin and streptomycin. Cells were grown at 37°C in a 5% CO_2 humidified incubator. Cells were grown to give 80%-90% confluence, harvested with Trypsin/EDTA, and washed with phosphate buffered saline (PBS).

Proteasome inhibitor treatment

In order to prevent protein degradation during cell lysis, the potent 20S proteasome inhibitor (PI) Bortezomib (PS-341) was used to treat the cells before carrying out cell lysis. PI Bortezomib was firstly dissolved in 1 ml Dimethyl sulfoxide (DMSO) to give a concentration of 13 mM and then diluted with DMEM/Glt/NCS to 50 µM before adding to cells. Aliquots of $1X10^8$ HEK293 cells grown in suspension were treated with proteasome inhibitor solution (50 µM) for 4 hours and washed three times with PBS. For HeLa cells, 50 µM PI treatment for 8 hours were performed before cell lysis.

Boiling water (500 µl) was applied directly into the frozen cell pellets and kept boiling for 15 minutes to destroy proteolytic activity. The sample was cooled to room temperature, sonicated on ice for 3 x 10 seconds at output level 4 with a 40% duty cycle (Branson Sonifier 250). Acetic acid was added to a final volume of 0.25% and centrifuged at 20,000xg for 20 minutes at 4°C to pellet insoluble material.

Polypeptide isolation

Molecular Weight Cut-Off (MWCO) + ERLIC (or SCX) approach

Reduction and alkylation

In order to collect low molecular weight proteins and peptides, 10 kDa, 30 kDa, and 50 kDa molecular weight cut-off (MWCO) filters (Amicon Ultra Kit) were used. An aliquot from the flow-through was quantified using a2D Quant Kit (Appendix 3) (Weist et al., 2008) to determine protein concentration. The top concentrated fraction from the MWCO filter was also collected and an aliquot quantified. Samples were brought to 5 mM dithiothreitol (DTT) and then incubated at 56°C for 30 minutes to reduce disulfide bonds. The sample reaction was cooled to room temperature before adding 55 mM iodoacetamide (IAA) solution to alkylate the free cys residues. The reaction was then incubated in the dark for 30 minutes followed by addition of more DTT (10 mM final concentration) to quench excess IAA.

Trypsin/ LysC digestion

Trypsin is a protease that is often used to cleave proteins to create a basic residue at the carboxyl terminus of the peptide for MS/MS analysis. Trypsin was mixed with 50 mM ammonium bicarbonate (ABC) to give 0.2 ng/µl before applying to samples. Trypsin was added at a ratio of 1:50 enzyme to protein (The ratio of 1:100 was attempted previously, but it was too low for peptide digestion). The reaction was incubated at 37°C for 16 hours. The digested sample then desalted using a C18 Toptip by washing with 3x150 µl 1% acetonitrile (ACN) and eluted with 2x150 µl 80% ACN/0.1% formic acid. The peptide mix was then dried in a SpeedVac at 45°C.

LysC has high specificity for lysine residues and creates larger peptides than trypsin. LysC was used as an alternative to digest proteins since some peptide sequences might contain too many

arginine residues, and those peptides would be cut into too small pieces (less than 6 amino acid residues) by trypsin. Procedures and the concentration of LysC used were the same as that performed with Trypsin digestion.

Polypeptide fraction by ERLIC

A PolyWax column (200 mm x 4.6 mm, 5 µm, 300 Å) was used when performing ERLIC using an Agilent Technologies 1100 Series HPLC combined with a degasser and automatic fraction collector. Flow rate was set at 0.8 ml/min with a wavelength of 254 nm. An aliquot of 1 mg of protein from the protein sample was loaded for fractionation (150 µl/load). Twenty-five fractions were collected over a 77-minute linear gradient beginning with a buffer solution containing 1% acetic acid in 90% acetonitrile (ACN) and ending with a buffer of 0.1% formic acid in 30% acetonitrile (ACN). Protein samples fractionated by ERLIC were not fractionated by SCX. Fractions were then evaporated on a SpeedVac before re-suspending in 15 µl 0.1% formic acid and then loaded onto a C18 column for LC-MS/MS analysis.

Polypeptide fraction by SCX

Samples were first desalted using a ZipTip prior to separation on an Agilent SCX column (50 mm x 4.6 mm, 5 µm, 300 Å) at a flow rate of 0.3ml/min beginning with buffer A containing 2% ACN/ 0.5% acetic acid followed by buffer B containing 2% ACN/0.5% acetic acid/ 250mM ammonium acetate, and running for 37 minutes on an Agilent 1100 chromatography system. An aliquot of less than 500 µg of protein from the protein sample was loaded for fractionation (150 µl/load). Fractions (250 µl) were collected in a microtitre plate and then pooled to give 9 fractions in total. These fractions were then evaporated on a SpeedVac followed by ZipTip cleaned up before re-suspending in 15 µl 0.1% formic acid for LC/MS/MS analysis.

SDS-PAGE gel LC-MS/MS approach

Tris-Tricine/Urea gels (SDS-PAGE)

Tris-Tricine/Urea gels (Separation gel: 16%, 6 M urea; stacker gel: 4%) were used for protein separation. Loading buffer was made with 150 mM Tris-HCl pH 7.0, 12% w/v SDS, 30% v/v glycerol, 6% v/v mercaptoethanol, 0.05% w/v Bromophenol Blue. Volume of loading buffer needs to be sufficient to keep ratio of SDS:neutral at least 10:1. Then, the loading buffer was mixed with protein samples prior to electrophoresis. A discontinuous buffer system consisted of anode buffer (1 M Tris-HCl, pH 8.9) and cathode buffer (1 M Tris-HCl, 1% SDS, pH 8.3) was used for electrophoresis. During electrophoresis, gels were initially run at 30 V and then switch to 200 V as samples reach the 4% stacker gel. Gels were then stained with Coomassie Blue G250 and imaged with Odyssey (LI-COR) (Preset: ProteinGel, Resolution: 169 μm, Channel: 700) prior to gel excision.

Reduction and alkylation

For visualization of gels, the protein sample from each lane corresponded to <20 kDa of the molecular weight (MW) marker were excised into several fractions between sizes15-20 kDa, 10-15 kDa, 3.5-10 kDa, and those less than 3.5 kDa. Gel slices of each respective fraction were diced into 1-2 mm pieces using a scalpel blade and were collected into 1.5 ml Eppendorf tubes for in-gel digestion prior to de-staining procedure. Gel pieces were then washed with ~500 µl of de-staining buffer (50% ACN, 50mM ABC). DTT (10 mM) was added to the gel pieces with 30 min-incubation at 56°C to reduce disulfide bonds. Samples were cooled to room temperature before adding a 55 mM iodoacetamide (IAA) solution. Samples were incubated in the dark at room temperature for 30 minutes. Dehydration of gel pieces was performed with 100% ACN followed by rehydration with Trypsin in 50 mM ABC. The digested sample then desalted using ZipTip on an Agilent C18 trap by

washing with 3x150 µl 1% ACN and eluted with 2x150 µl 80% ACN/0.1% formic acid. The peptide mix was then dried in a SpeedVac at 45°C.

Peptide extraction

To extract peptides from the gel pieces, extraction buffer (50% ACN/ 0.1% TFA) was added to samples with sonication for 10 minutes. Samples were then evaporated in a SpeedVac at 45°C. Before LC-MS/MS analysis, samples were desalted using ZipTip by washing with 3x10 ul 1% ACN and eluted with 10 ul 80% ACN/0.1% TFA.

Nano flow LC-MS/MS analysis

An electrospray quadrapole- time-of-flight mass spectrometer (ESI-Q-TOF MS) (Applied Biosystems TripleTOF 5600 System) fitted with Nanospray III source was used with across m/z 350 – 1800. The top 20 multiply charged (2+ to 5+) peptides that had more than 100 counts intensity were picked for fragmentation by collision-induced dissociation (CID) across m/z 40 - 1800 for 0.05 sec, followed by exclusion for 30 sec after 2 occurrences, with up to 20 MS/MS experiments per cycle (0.05 sec / spectrum).

Database search for protein analysis

Raw MS data files were first converted to MASCOT Generic Format using the mgf processing script (version 1.3) accessed via Peakview version 1.2 (ABSciex). UniProKB/Swiss-Prot database was used for analysis of the mass spectra of protein samples using MASCOT search engine from Matrix Science, accessed via the Australian Proteomics Computational Facility. The parameters for MASCOT MS/MS search are listed in Table A4.

Database search for sPEP identification

Identification of sPEPs from the protein mix was performed using MASCOT search engine. Details of the process are described in Figure 2.1. According to determine how good the data searched against to the RefSeq database. Any peptide spectrum matches (PSMs) with a score less than 40 were discarded since the likelihood that of a false positive increases as the score decreases. The remaining MS/MS data matched in human RNA database were examined to confirm that they met the criteria for a sPEP. Those MS/MS spectra that match main ORFs in SwissProt were excluded and everything else was re-searched in the RefSeq database. A good match must have high score obtained by MASCOT algorithm with low error values, and well-matched MS/MS spectrum in b- and/or y-ion coverage. In addition, the predicted peptide sequence should not sit in the mCDS of a gene, determined by conducting tBLASTn (NCBI) searches. In addition to MASCOT searches, the MS/MS data was also searched against our inhouse sORF list, and as well as to the Human alternative open reading frame (HaltORF) database (Vanderperre et al., 2012), using ProteinPilot^{TM} (ABSCIEX), which is able to utilize a different search algorithm and it enables simultaneous searches that contained multiple peptide modifications, which is a limitation in the MASCOT MS/MS ion search algorithm.

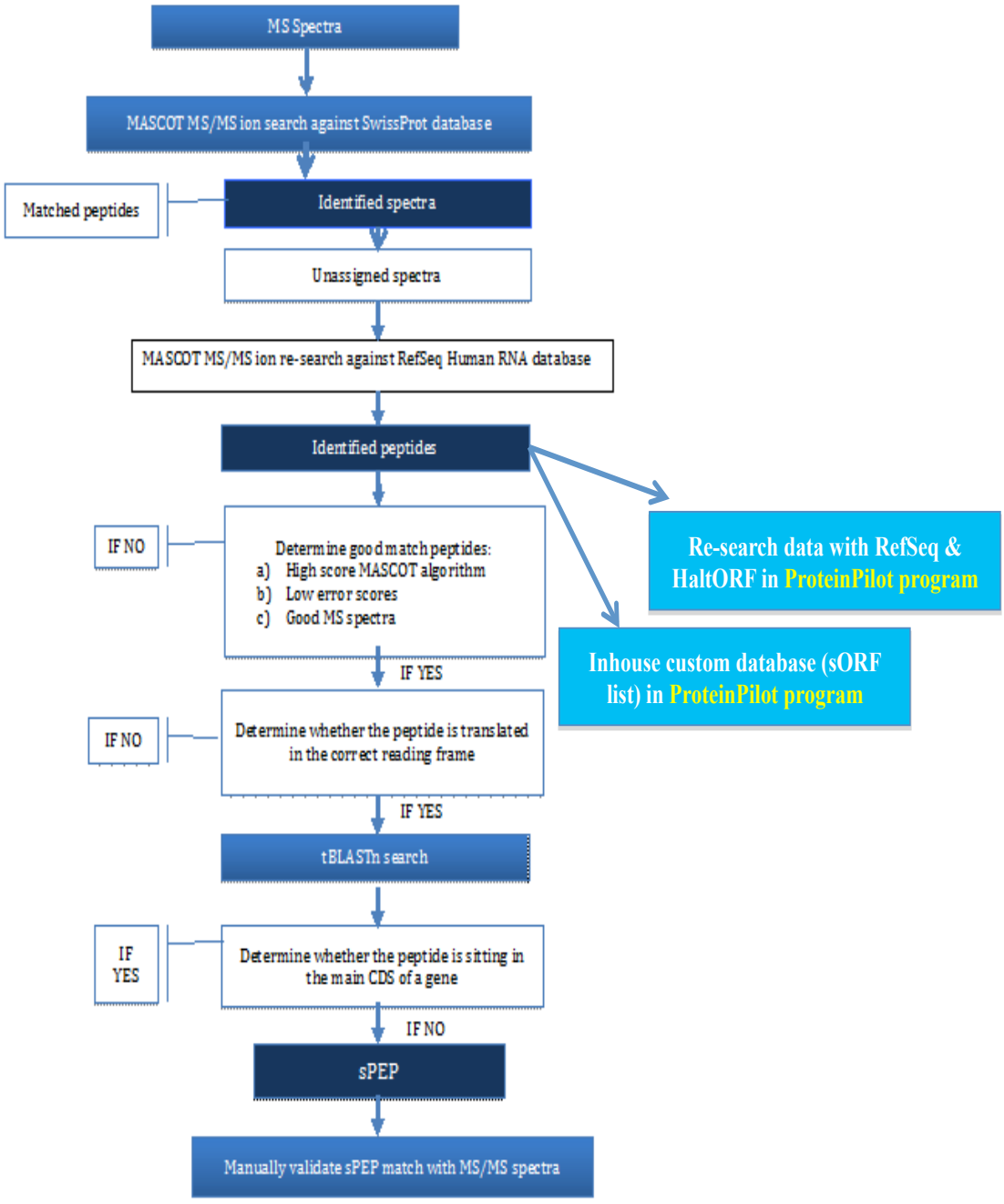


Figure 2.1. Flowchart indicating the process used to identify sPEPs from MS/MS data using the MASCOT and ProteinPilot™ searches.

Chapter 3

Results of proteomic approaches

Introduction

In order to reduce sample complexity in peptide enrichments,, low molecular weight proteins from HeLa and HEK293 cell lysates were extracted by either SDS-PAGE or ERLIC fractionation. These extracts were digested with trypsin or LysC and analysed by nano LC-MS/MS. The process to identifying sPEPs from the protein mix was performed using MASCOT search engine. In addition to MASCOT searches, the MS/MS data was also searched against our inhouse sORF list, and as well as to the Human alternative open reading frame (HaltORF) database (Vanderperre et al., 2012), using ProteinPilot^{TM} (ABSCIEX), which is able to utilize a different search algorithm and it enables simultaneous searches that contained multiple peptide modifications, which is a limitation in the MASCOT MS/MS ion search algorithm.

The proteomic approaches used in my project were analysed by comparing the resulting data from each critical procedure. In order to prevent protein degradation during cell lysis, the potent 20S proteasome inhibitor (PI) Bortezomib (PS-341) was used to treat the cells before carrying out cell lysis. Different cell batches with/without PI treatment were recorded and analysed by measuring the amount of protein using the 2D quant kit. The amount of protein products obtained from ERLIC and SCX in different rounds of attempts were analysed and compared from MASCOT searches. In addition, the efficiency of these peptide enrichment strategies was also analysed by comparing the identification of protein products after LC-MS/MS analysis.

Proteasome inhibitor treatment of cells

Previous work in the laboratory had shown that half life of exogenous sPEPs was increased by the addition of a proteasome inhibitor (Andrews 2012). In order to prevent protein degradation, the potent 20S proteasome inhibitor (PI) Bortezomib (PS-341) was used to treat the cells before cell lysis. A test round of HeLa and HEK293 cells treated with PI in different incubation times (1 hour,

4 hours, 8 hours) was performed (Figure 3.1&3.2) to estimate the incubation time of PI treatment. HeLa cells treated with 50 μM for 8 hr. HeLa cells were still healthy after 8-hour 50 μM PI treatment. Therefore, HeLa cells treated with 8-hour 50 μM PI treatment were used for subsequent experiments. For HEK293 cells, cells were still healthy after 4-hour 50 μM PI treatment but cells began to detach and float on the flask after 8-hour 50 μM PI treatment. Therefore, 4-hour 50 μM PI treatment was demonstrated in HEK293 cells.

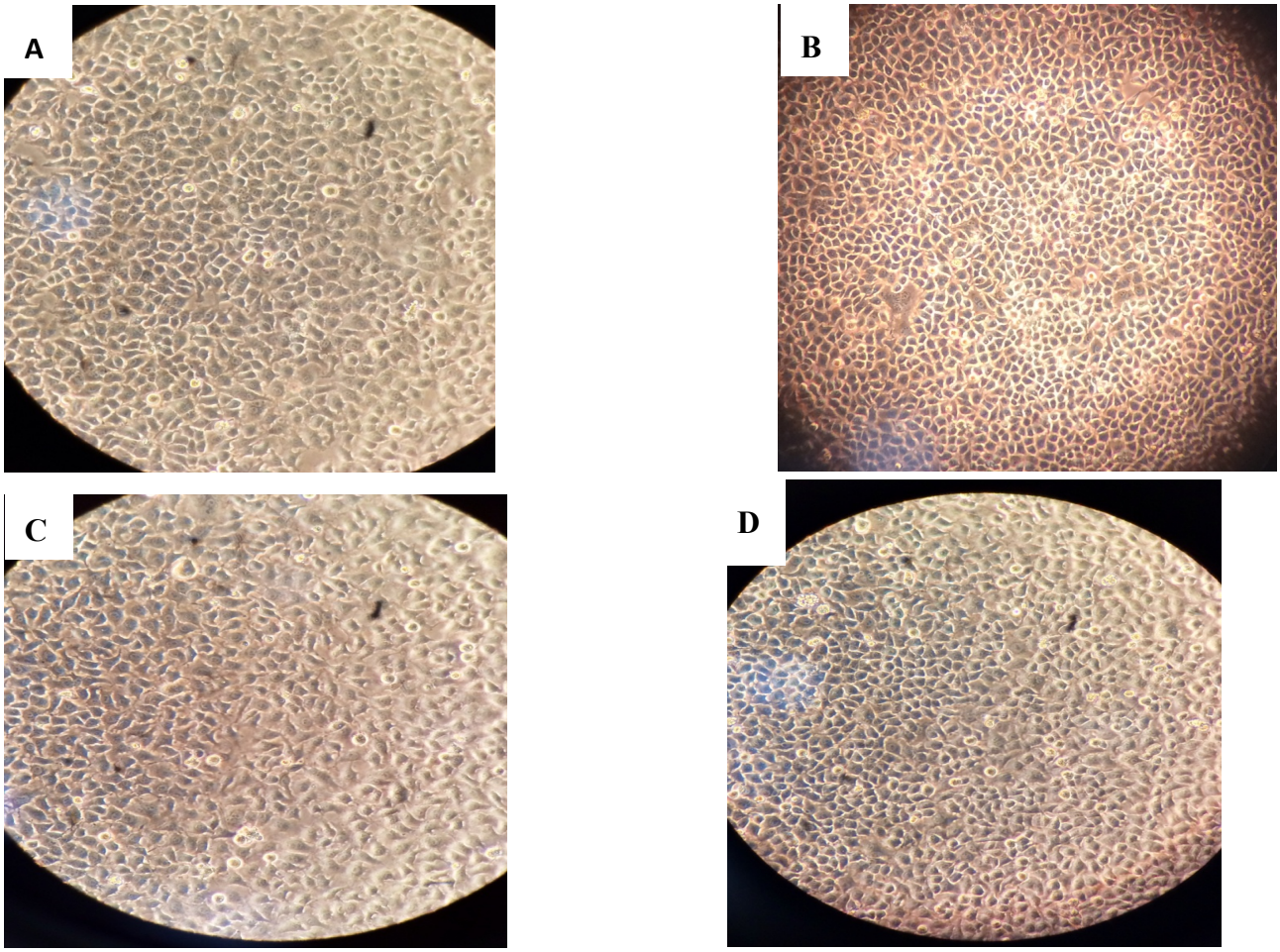


Figure 3.1. Results of PI treatment in HeLa cells with different incubation times.
A) Confluence of HeLa cells before PI treatment: ~70%. 10 x 10 magnification. B) HeLa cells treated with 50 μM for 1 hr.10 x 10 magnification. C) HeLa cells treated with 50 μM for 4 hr. 10 x 10 magnification. D) HeLa cells treated with 50 μM for 8 hr. HeLa cells were still healthy after 8-hour 50 μM PI treatment.10 x 10 magnification. Therefore, HeLa cells treated with 8-hour 50 μM PI treatment were used for subsequent experiments.

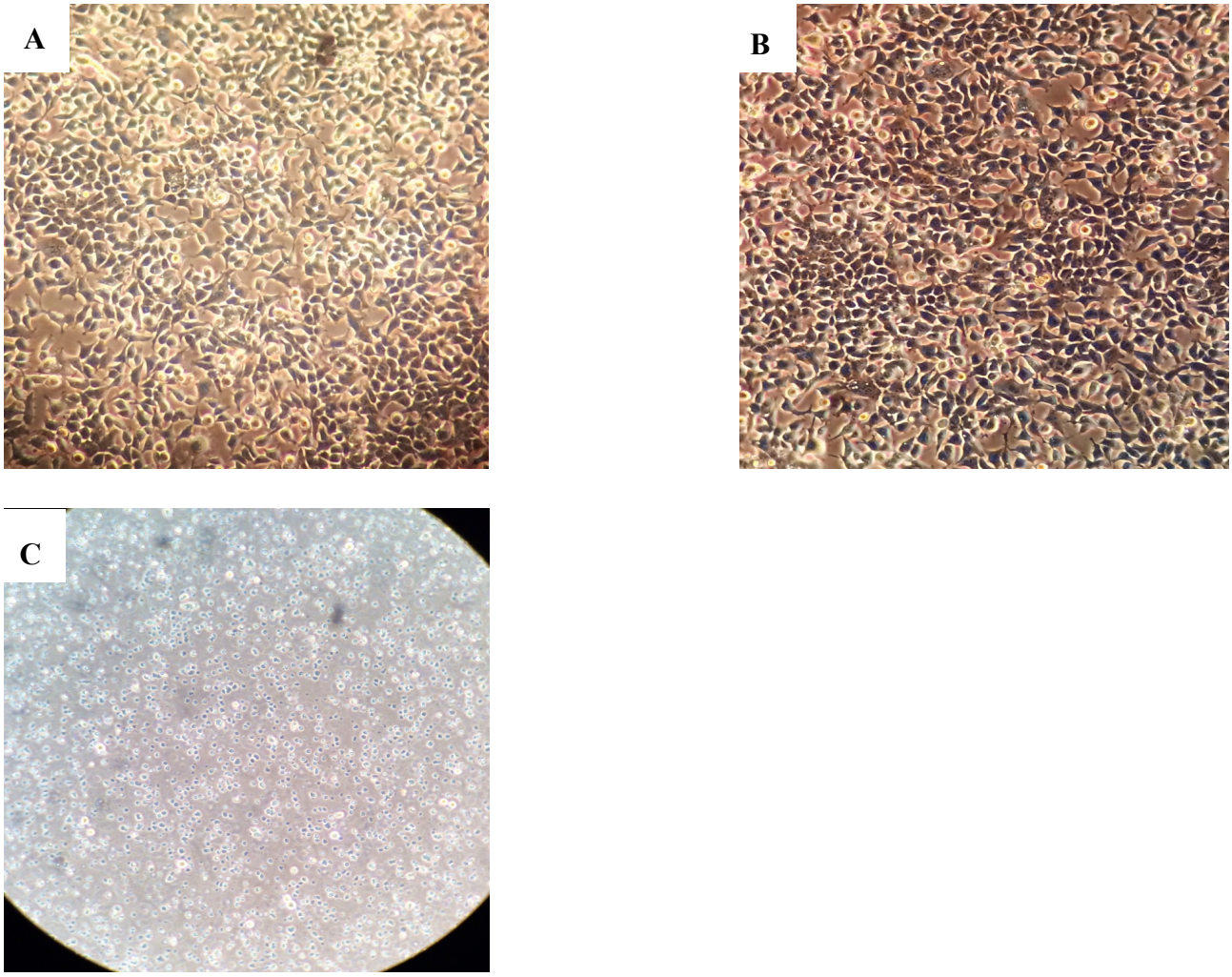


Figure 3.2. Results of PI treatment in HEK293 cells with different incubation times.
A) Confluence of HEK293 cells before PI treatment: ~70%.10 x 10 magnification. B) HEK293 cells treated with 50 µM for 4 hr. 10 x 10 magnification. C) HEK293 cells treated with 50 µM for 8 hr. HEK293 cells were detached and floating on the flask after 8-hour 50 µM PI treatment. 10 x 10 magnification. Therefore, HEK293 cells treated with 4-hour 50 µM PI treatment were used for subsequent experiments.

Different cell batches with/without PI treatment were analysed by measuring the amount of protein using 2D quant kit (Figure 3.3). For HeLa cells, ~80% more protein was obtained from the cells with 50 µM PI treatment compared to those without PI treatment (Figure 3.4). The comparison of HEK293 cells with/without PI treatment indicated that the amount of protein yield was doubled

from cells with 50 µM PI treatment according to the results from 2D quant measurement (Figure 3.5). However, in MASCOT searches, no significant increase was detected in the numbers of peptide yield from cells with PI treatment (Figure 3.6).

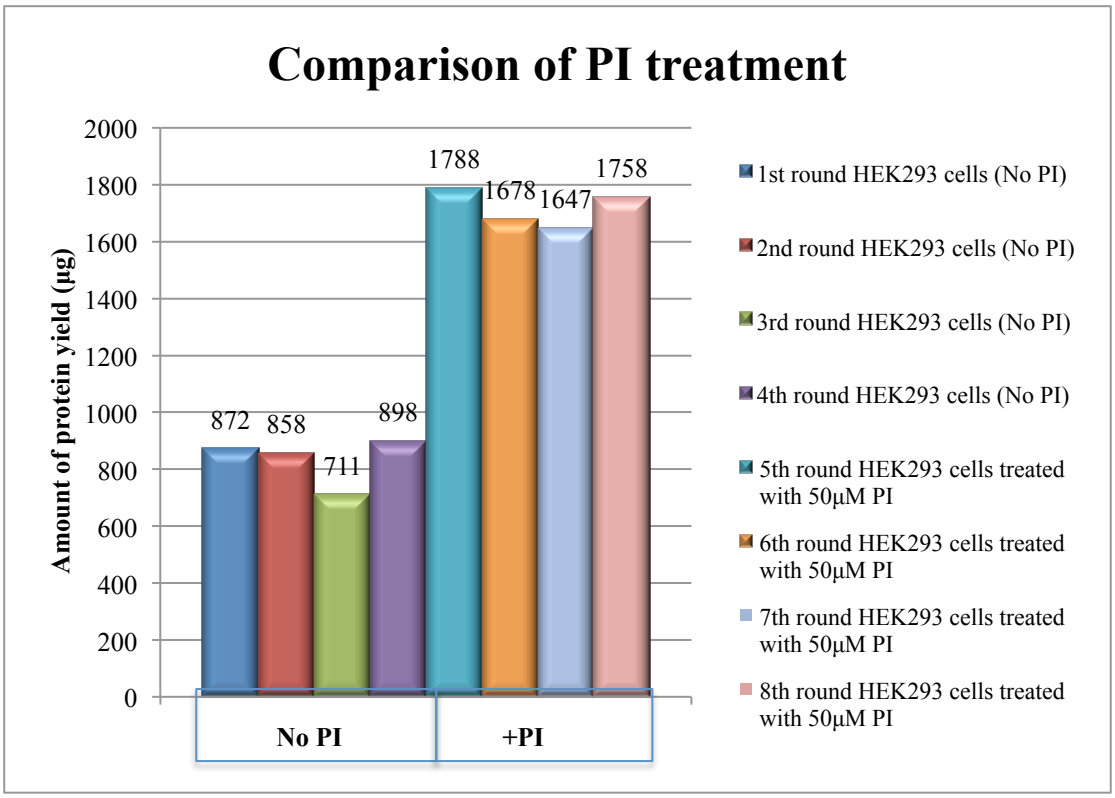


Figure 3.3. Amounts of protein yields from different batches of HEK293 cell lysates with/without proteasome inhibitor treatment

Different HEK293 cell batches with/without PI treatment were analysed by measuring the amount of protein yield using 2D quant kit. The first four lanes in the figure show protein yield from cells without PI treatment while the following four lanes indicates protein yield from cells treated with PI. The number at the top of each bar indicate actual yield in µg. Each bar represents a separate experiment.

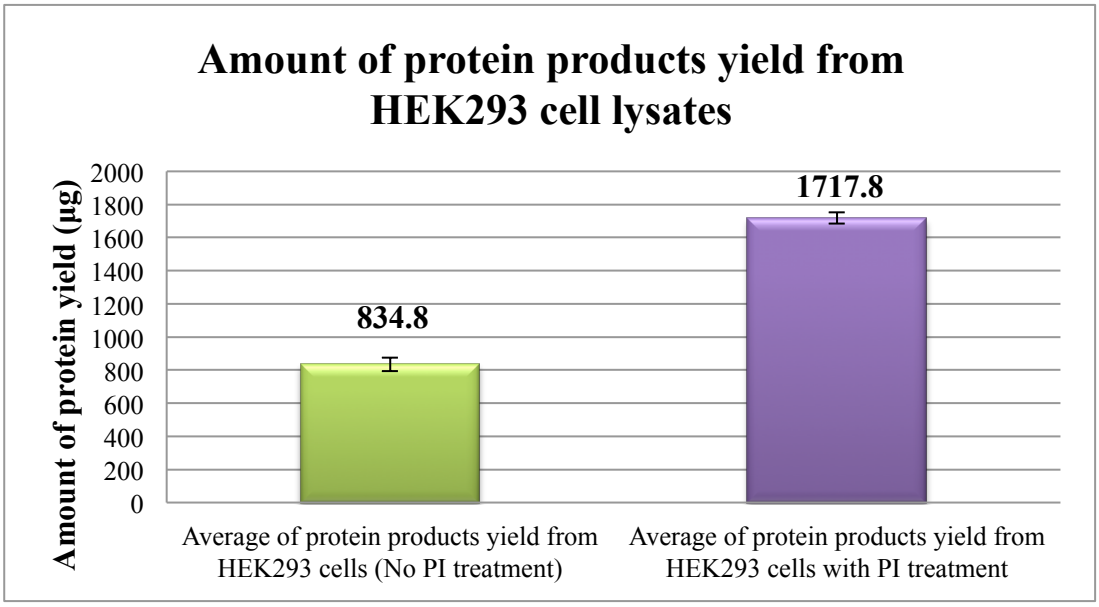


Figure 3.4. Amount of protein products yield from HEK293 cell lysates

A comparison of HEK293 cells with/without PI treatment indicated that the amount of protein yield was doubled from the cells treated with 50 µM PI. The average value is based on 8 distinct experiments with individual HEK293 cell batch. Standard error of the mean was displayed in the error bar.

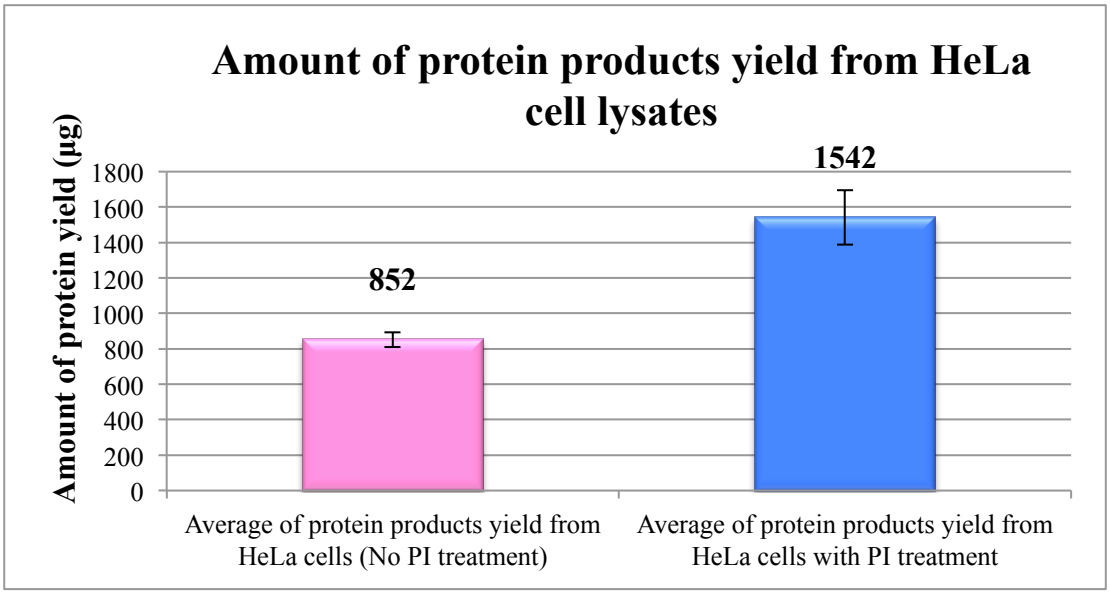


Figure 3.5. Amount of protein products yield from HeLa cell lysates

Results indicate ~80% more proteins yielded from the cells with 50 µM PI treatment than those without PI treatment from HeLa cell lysates. The average value is based on 4 distinct experiments with individual HeLa cell batch. Standard error of the mean was displayed in the error bar.

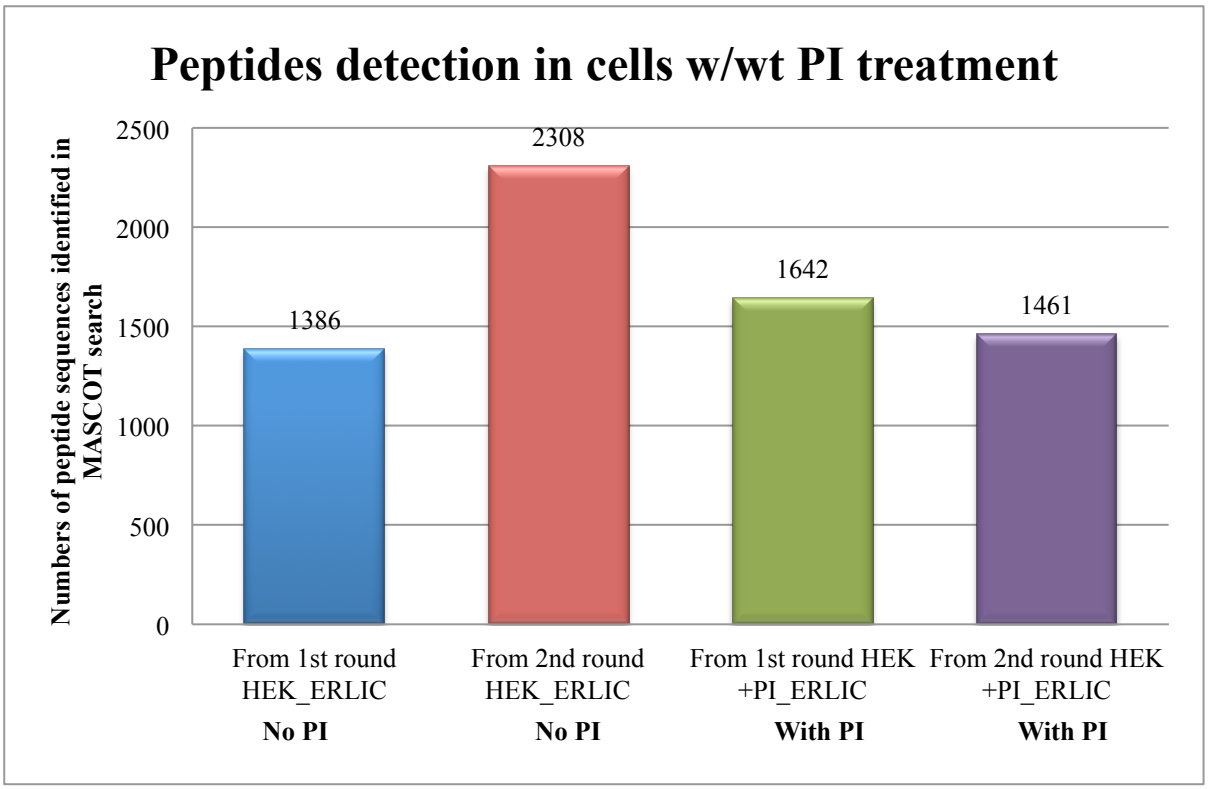


Figure 3.6. Numbers of peptide detected in cells with/without PI from MASCOT search

According to MASCOT search results, the numbers of peptide did not have detectable increase in the cells treated with PI.

Analysis of peptide enrichment strategies

Molecular Weight Cut-Off (MWCO) + ERLIC (or SCX) approach

Both ERLIC and SCX have been shown to increase detection sensitivity in peptide separation compared to hydrophilic interaction liquid chromatography (HILIC) (Zarei et al., 2011). Although the ERLIC approach in enriching peptides for MS-based identification has been reported to provide more advantages than the SCX approach (Gan et al., 2008), SCX has been reported to have higher identification of phosphopeptides (Zarei et al., 2011). In addition, ERLIC is shown to have better result in the separation of multi-phosphorylated peptides, while SCX is suited for the fractionation of mono-phosphorylated peptides (Zarei et al., 2012). Therefore, both approaches were performed

in my project to improve the separation and potential discovery of new sPEPs. The fractionated protein products obtained from ERLIC and SCX from a number of separate experiments was analysed by LC-MS/MS followed by MASCOT searching (Figure 3.7). From the results of protein identification of sPEPs from ERLIC and SCX, the last ERLIC experiment had approximate twice as many protein identifications as the first ERLIC experiment (Figure 3.7). As a batch of HEK293 cells treated with PI was used for lysis, followed by a 30 kDa molecular weight cut-off (MWCO) filter for collection of small proteins in the last ERLIC experiment, while a batch of HEK293 cells was used without PI treatment, followed by a 10 kDa MWCO filter. Similar results were detected in the SCX experiments. The total number of proteins identified in each experiment varied quite a lot and may be due to differences in each protocol, including cell culturing (PI treatment) and peptide separation and extraction strategies. From the results, the number of protein products obtained via ERLIC fractionation is more than that via SCX when 1 mg of protein was used as the starting material for both approaches. The average value is based on 10 distinct experiments. The total number of sPEP identifications from ERLIC and SCX was calculated for comparison (Figure 3.8). Protein product identification from HEK293 cell lysates fractionated via ERLIC was ~50% higher than that via SCX. However, the difference in the protein product identification between ERLIC and SCX may be due to the different numbers of fractions collected from ERLIC (25 fractions) and SCX (10 fractions), and as well as the fact the total amount of starting materials that went on each column were different. During analysis of MS/MS data, some overlaps of different subset of peptides were observed in samples from ERLIC and SCX approaches.

SDS-PAGE gel LC-MS/MS approach

Tris-Tricine/Urea gels have been commonly used for protein separation in the mass range 1-100 kDa, especially for the resolution of proteins smaller than 30 kDa (Schägger, 2006). 10% acrylamide gels have been shown to have rapid separation and relatively wide mass range coverage

(2-100 kDa) (Schägger, 2006). As the resolution power of Tris-Tricine/Urea gels increases for small proteins with increased acrylamide concentration, 16% acrylamide gels were used in these experiments. Since excess urea can reduce the electrophoretic mobility, and as well as cause oligomerisation in membrane proteins, the concentration of urea under 6 M is recommended (Schägger, 2006). Therefore, 6 M of urea was used in my experiment. After separation of HEK293 or HeLa proteins on Tris-Tricine SDS-PAGE gels the gels were then stained with Coomassie Blue G250 to detect protein bands. By using 16% acrylamide gels, excellent separation of protein <20 kDa was obtained as illustrated in Figure 3.9. In order to increase total protein available for LC-MS/MS, multiple lanes of protein lysates were run in gels.

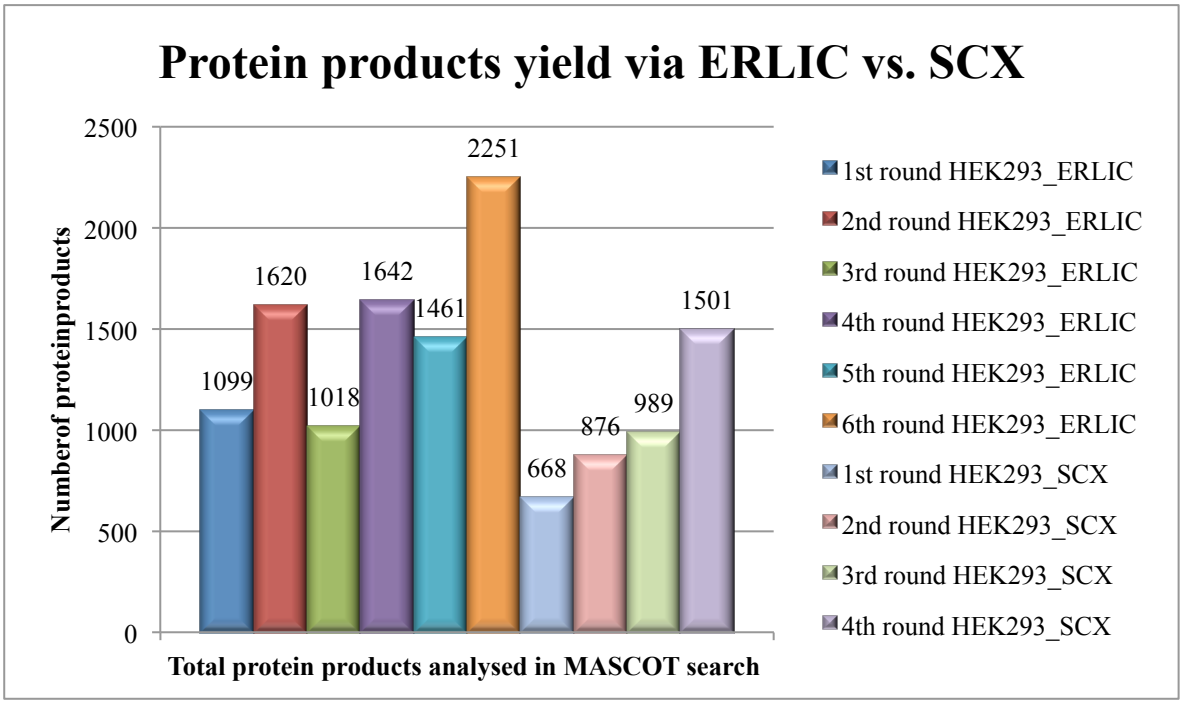


Figure 3.7. Protein products identification via ERLIC and SCX

The number of protein products obtained from HEK293 cell lysates via ERLIC and SCX fractionation in different rounds of attempts were calculated. From the results, the number of protein products obtained via ERLIC fractionation is relatively more than that via SCX when 1 mg of protein was used as the starting material for both approaches. The average value is based on 10 distinct experiments.

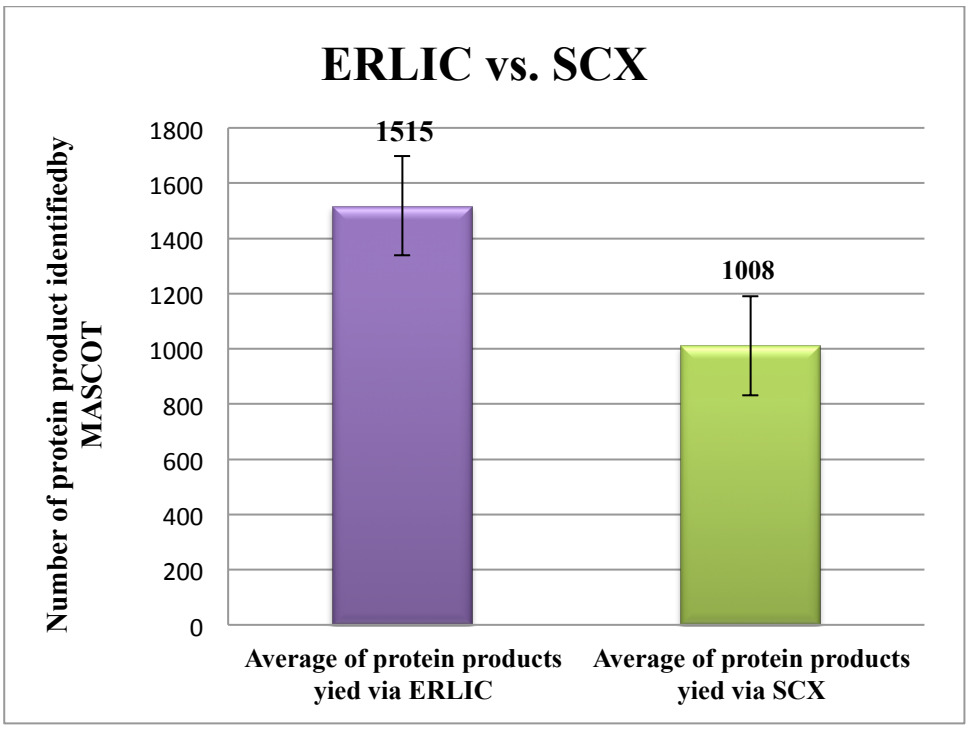


Figure 3.8. Comparison of protein products identification via ERLIC and SCX
Protein products identification from HEK293 cell lysates fractionated via ERLIC was ~50% higher than that via SCX. The average value is based on 10 distinct trials through ERLIC and SCX. Standard error of the mean was displayed in the error bar.

The efficiency of these peptide enrichment strategies performed in my experiments was analysed by comparing the identification of protein products after LC-MS/MS analysis. From the resulting MS/MS data, 1515 protein products on average yielded from ERLIC approach while only 885 protein products on average were obtained from the SDS-PAGE gel approach (Figure 3.10). In the comparison between SCX and SDS-PAGE approach, 1008 protein products in average yielded from SCX approach while only 885 protein products in average yielded from SDS-PAGE gel approach (Figure 3.11). Overall, both ERLIC and SCX approaches resulted more protein products from the same amount of starting material than that from SDS-PAGE gel approach. In addition, in my MS/MS data, more sPEPs were identified from ERLIC and SCX approaches than SDS-PAGE gel approach.

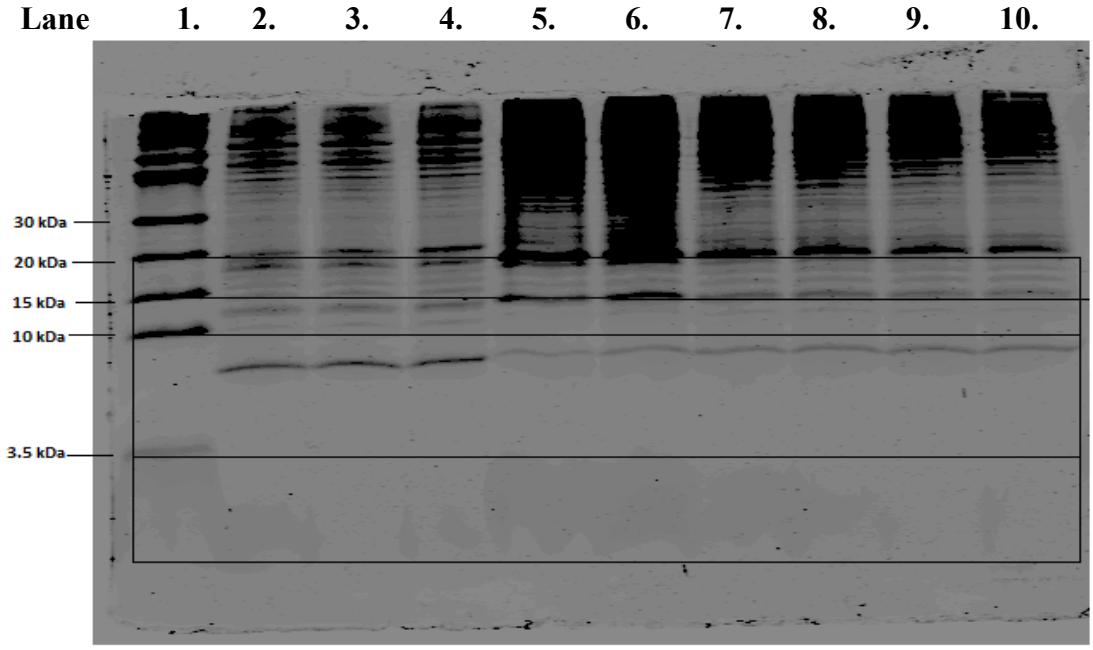


Figure 3.9. Results of SDS-PAGE gel electrophoresis in HEK293 cell lysates

Proteins that less than 20 kDa molecular weights were excised from the gels in four sections: 15-20 kDa, 10-15 kDa, 3.5-10 kDa, and those less than 3.5 kDa as illustrated. Gel pieces were diced into 1-2 mm pieces and transferred into 1.5 ml Eppendorf tubes, followed by de-staining with 50% ACN/ 50 mM ABC to remove Coomassie Blue.

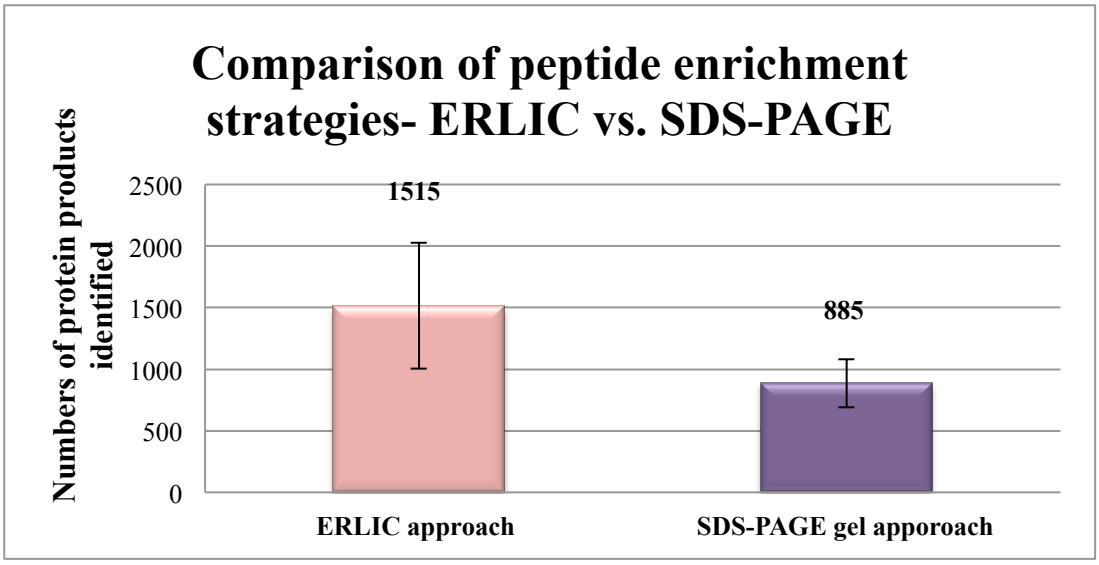


Figure 3.10. Analysis of peptide enrichment strategies- ERLIC vs. SDS-PAGE

The efficiency of these peptide enrichment strategies performed in my experiments was analysed by comparing the identification of protein products after LC-MS/MS analysis. From the resulting MS/MS data, 1515 protein products in average yielded from ERLIC approach while only 885 protein products in average yielded from SDS-PAGE gel approach. Standard error of the mean was displayed in the error bar.

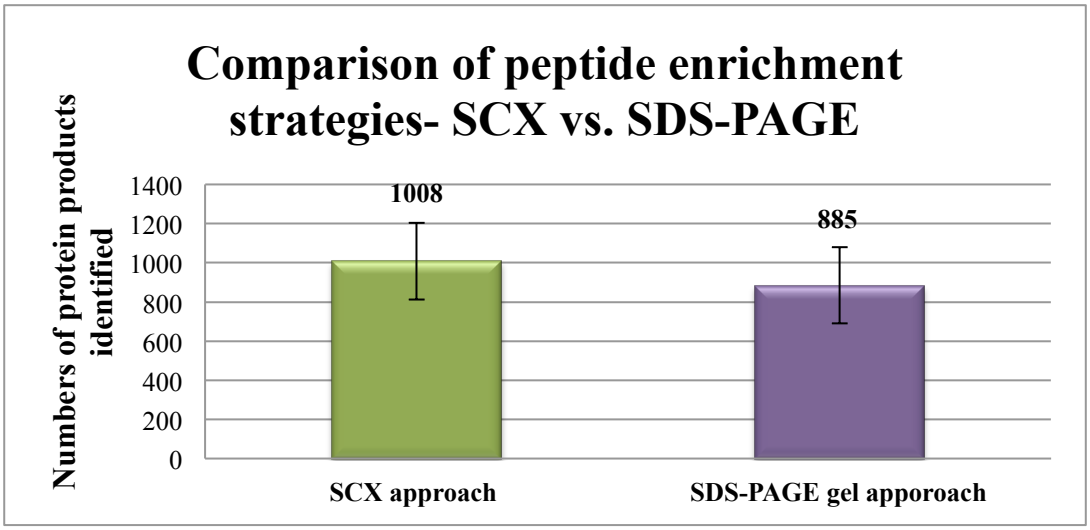


Figure 3.11. Analysis of peptide enrichment strategies- SCX vs. SDS-PAGE

The efficiency of these peptide enrichment strategies performed in my experiments was analysed by comparing the identification of protein products after LC-MS/MS analysis. From the resulting MS/MS data, 1008 protein products in average yielded from SCX approach while only 885 protein products in average yielded from SDS-PAGE gel approach. Standard error of the mean was displayed in the error bar.

MS analysis

LC-MS/MS proteomics had been reported to enrich small polypeptides (Tinoco et al., 2010). Since the optimal size of polypeptides for LC-MS/MS detection is approximately 10 to 20 amino acids, trypsin digest has been reported (Slavoff, Mitchell et al. 2013) as a critical step for high sensitivity peptide detection because the average size of the tryptic peptides generated are about the right size (less than 50 amino acids in length) for MS analysis.

To identify sPEPs, first, the resulting MS/MS data was searched against the SwissProt using MASCOT to filter out known proteins, and all unmatched spectra were further searched against the human RefSeq database. Any peptide spectrum matches (PSMs) with a score less than 40 were discarded since the likelihood that of a false positive increases as the score decreases. The remaining MS/MS data matched in human RNA database were examined manually to confirm that they met the criteria for a sPEP. The set of criteria includes: 1) a sequence tag of five consecutive b or y ions, 2) a precursor mass error of <50 ppm, 3) excellent sequence coverage. Using these criteria, nearly 50% of the remaining peptides were discarded. Peptides that met the criteria were then searched to determine whether they are translated in the correct reading frame by performing. tBLASTn searches for these translated sPEPs to determine their location in the corresponding gene sequence. Only those sPEPs positioned outside and / or out of the main CDS were collected for further analysis. After these examinations, approximately 1% of the remaining sPEPs remained from the candidate set. These MS/MS spectra were validated manually for the final confirmation of sPEP identification.

To identify sPEPs missed by MASCOT and also to check for post-translational modifications (PTMs) which is limited in MASCOT searches due to the algorithm used, those MS/MS data filtered after SwissProt searches were also analysed against RefSeq Human RNA database and

cross-checked against HaltORF database (Vanderperre et al., 2012) using ProteinPilot^{TM} $v4.5$. Lastly, to determine if sPEPs had been identified previously in proteomic studies, sORFs were analysed against our in-house sORFs list, which contains all sORFs identified up to date.

In this project, 1,879,681 MS/MS spectra were obtained in my experiments. The MS/MS data matched in human RNA database were examined to confirm that they met the criteria for a sPEP. Those MS/MS spectra that match main ORFs in SwissProt were excluded and everything else was re-searched in the RefSeq database, and this gave out more than 12,000 protein products (ORF hits) in the resulting pool. Any peptide spectrum matches (PSMs) with a score less than 40 were discarded, and it gave out more than 6,000 sORF after the analysis (Table 3.1). To identify those protein products, the sORFs were examined by hand according to the high score obtained in MASCOT algorithm with low error values, and well-matched MS/MS spectrum in b- and/or y-ion coverage. In addition, the predicted peptide sequence should not sit in the mCDS of a gene, determined by conducting tBLASTn (NCBI) searches. From the protein products analysed in MASCOT, eight sORFs have been identified and 11 distinct sPEPs matching back to those eight sORFs have been observed. Three of these have been confirmed as novel sPEPs (Table 3.2). From the searches in ProteinPilot^{TM} $v4.5$, two sPEPs were identified and which were also found in MASCOT searches. The MS/MS raw data of those sPEPs are shown in Figure 3.12. Among these sPEPs, five sPEPs are encoded from uORFs, two sPEPs are from ncRNAs, and four sPEPs are encoded from oORFs. Results showed that some of those 11 sPEPs have appeared reproducibly in different cell batches throughout the experiments. This result increases the confidence of sPEP confirmation and identification.

	Number
Total number of MS/MS spectra (SwissProt)	1,879,681
Forwarded to the Human_RNA database	1,853,964
Protein products matched after SwissProt search	13,928 (includes duplicates)
Protein products matched after Human_RNA search	12,028 (includes duplicates)
Peptide sequences matched after MASCOT score >40	6,204 (includes duplicates)
sORFs identified through MASCOT search	8 (NM_019048.2; NR_003608.1; NM_015532.3; NM_080670.2; NM_004540.3; NR_024006.1; NM_020123.3; NM_007039.3)
sPEPs identified through MASCOT search	11 (NM_019048.2; NR_003608.1; NM_015532.3; NM_080670.2; NM_004540.3; NR_024006.1; NM_020123.3; NM_007039.3)
Protein products found through ProteinPilot™	19,440(includes duplicates)
sPEPs identified through ProteinPilot™	2 (NM_019048.2)

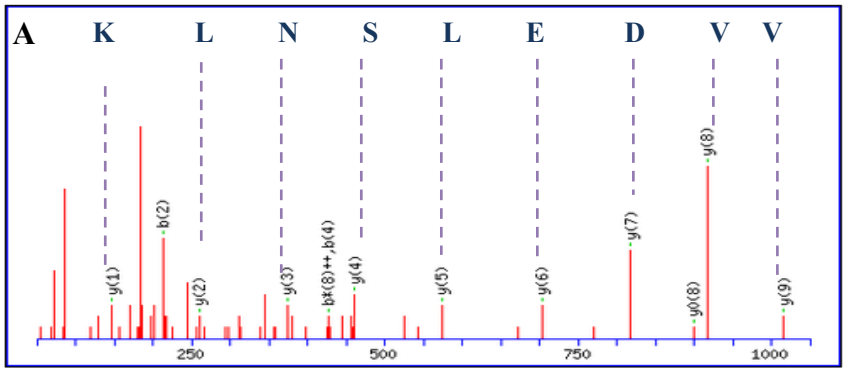
Table 3.1. Supplementary data of MS/MS results.

Gene name (NCBI)	**Peptide Sequence** (Actual peptides identified in this project are highlighted in red)
NM_019048.2 Homo sapiens asparagine synthetase domain	MPSRGTRPEDSSVLIPTDNSTPHKEDLSSKIKEQKIVVDELSNLKKNRKV

containing 1 (*ASNSD1*),mRNA	YRQQQNSNIFFLADRTEMLSESKNILDELKKEYQEIENLDKTKIKK (Also identified by(Slavoff et al., 2013, Oyama et al., 2007)
NM_019048.2 Homo sapiens asparagine synthetase domain containing 1 (*ASNSD1*),mRNA	MPSRGTRPEDSSVLIPTDNSTPHKEDLSSKIKEQKIVVDELSNLKKNRKV YRQQQNSNIFFLADRTEMLSESKNILDELKKEYQEIENLDKTKIKK (Also identified by(Slavoff et al., 2013, Oyama et al., 2007)
NM_019048.2 Homo sapiens asparagine synthetase domain containing 1 (*ASNSD1*),mRNA	MPSRGTRPEDSSVLIPTDNSTPHKEDLSSKIKEQKIVVDELSNLKKNRKV YRQQQNSNIFFLADRTEMLSESKNILDELKKEYQEIENLDKTKIKK (Also identified by (Slavoff et al., 2013, Oyama et al., 2007)
NM_019048.2 Homo sapiens asparagine synthetase domain containing 1 (*ASNSD1*),mRNA	MPSRGTRPEDSSVLIPTDNSTPHKEDLSSKIKEQKIVVDELSNLKKNRKV YRQQQNSNIFFLADRTEMLSESKNILDELKKEYQEIENLDKTKIKK (Also identified by(Slavoff et al., 2013, Oyama et al., 2007)
NR_003608.1 Homo sapiens tubulin, alpha 3f, pseudogene (*TUBA3FP*), non-coding RNA	MSGSCQRSGEDKKQEEEATAACGRLA GVPEAKQGPKADSDSDLETEGARGRGQARLLPLGASPAGVVGGGLAPP RRQETSVQQGT (Also identified by(Slavoff et al., 2013)
NM_015532.3 Homo sapiens polymerase (RNA) II (DNA directed) polypeptide M (*POLR2M*), transcript variant 1, mRNA	MATPARAPESPPSADPALVAGPAEEAECPPPRQPQPAQNVLAAPRLRAP SSRGLGAAEFGGAAGNVEAPGETFAQRKIHLQIARQR (Identified by (Oyama et al., 2007)
NM_080670.2 Homo sapiens solute carrier family 35, member A4 (*SLC35A4*), mRNA	MADDKDSLPKLKDLAFLKNQLESLQRRVEDEVNSGVGQDGSLLSSPFL KGFLAGYVVAKLRASAVLGFAVGTCTGIYAAQAYAVPNVEKTLRDYL QLLRKGPD (Identified by Wilson Ng, our research group 2012)
NM_004540.3 Homo sapiens neural cell adhesion molecule 2 (*NCAM2*), mRNA	MTVKLQAELEGIKRACTLILNMPPSLYQTKQFITLGKEILSI (Identified by (Vanderperre et al., 2013)

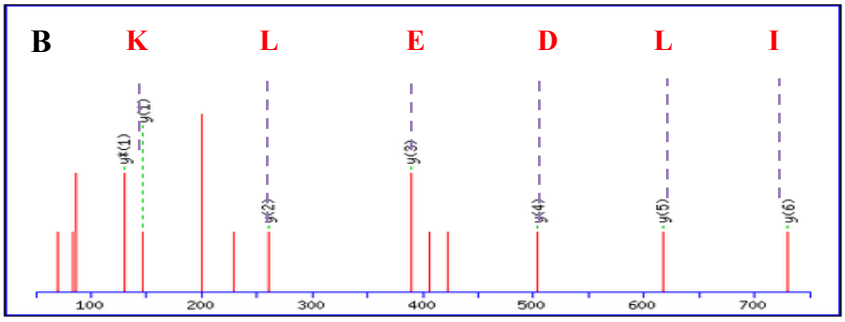
NR_024006.1 Homo sapiens uncharacterized LOC92973 (*FP588*) (*LINC000950*), non-coding RNA	ISNGSDEISLP (Novel peptide)
NM_020123.3 Homo sapiens transmembrane 9 superfamily member 3 (*TM9SF3*), mRNA	ATAAAEEAAAGPGPVR (Novel peptide)
NM_007039.3 Homo sapiens protein tyrosine phosphatase, non-receptor type 21 (*PTPN21*), mRNA	MGLYCHTGTITEYLSRFLADGMGTGNCNYSNGDSRRGGWKGEEL (Novel peptide)

Table 3.2. 11 sPEPs identified from proteomic and peptidomic process in this project.



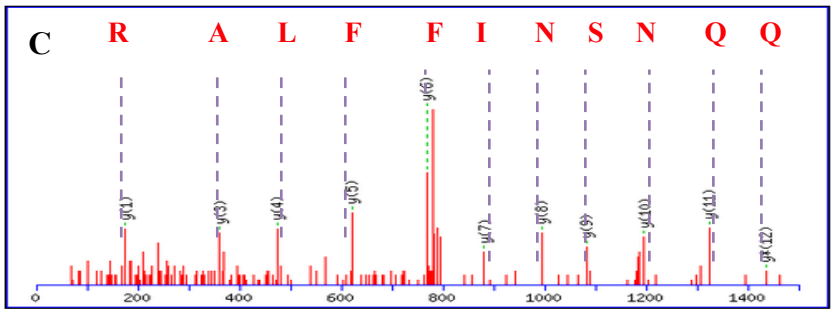
#	b	b^{++}	b*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	114.0913	57.5493					I							10
2	213.1598	107.0835					V	1016.5623	508.7848	999.5357	500.2715	998.5517	499.7795	9
3	312.2282	156.6177					V	917.4938	459.2506	900.4673	450.7373	899.4833	450.2453	8
4	427.2551	214.1312			409.2445	205.1259	D	818.4254	409.7163	801.3989	401.2031	800.4149	400.7111	7
5	556.2977	278.6525			538.2871	269.6472	E	703.3985	352.2029	686.3719	343.6896	685.3879	343.1976	6
6	669.3818	335.1945			651.3712	326.1892	L	574.3559	287.6816	557.3293	279.1683	556.3453	278.6763	5
7	756.4138	378.7105			738.4032	369.7053	S	461.2718	231.1395	444.2453	222.6263	443.2613	222.1343	4
8	870.4567	435.7320	853.4302	427.2187	852.4462	426.7267	N	374.2398	187.6235	357.2132	179.1103			3
9	983.5408	492.2740	966.5142	483.7608	965.5302	483.2687	L	260.1969	130.6021	243.1703	122.0888			2
10							K	147.1128	74.0600	130.0863	65.5468			1

A) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide IVVDELSNLK, found in the uORF of *ASNSD1* mRNA sequence.



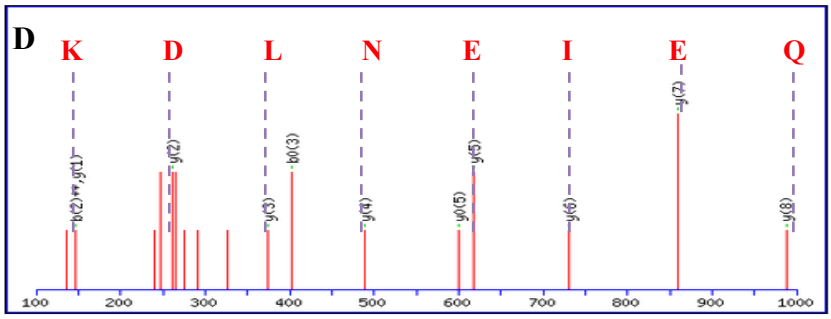
#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	115.0502	58.0287	98.0237	49.5155			N							7
2	228.1343	114.5708	211.1077	106.0575			I	*730.4345*	365.7209	713.4080	357.2076	712.4240	356.7156	6
3	341.2183	171.1128	324.1918	162.5995			L	*617.3505*	309.1789	600.3239	300.6656	599.3399	300.1736	5
4	456.2453	228.6263	439.2187	220.1130	438.2347	219.6210	D	*504.2664*	252.6368	487.2399	244.1236	486.2558	243.6316	4
5	585.2879	293.1476	568.2613	284.6343	567.2773	284.1423	E	*389.2395*	195.1234	372.2129	186.6101	371.2289	186.1181	3
6	698.3719	349.6896	681.3454	341.1763	680.3614	340.6843	L	*260.1969*	130.6021	243.1703	122.0888			2
7							K	*147.1128*	74.0600	130.0863	65.5468			1

B) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide NILDELK, found in the uORF of *ASNSD1* mRNA sequence.



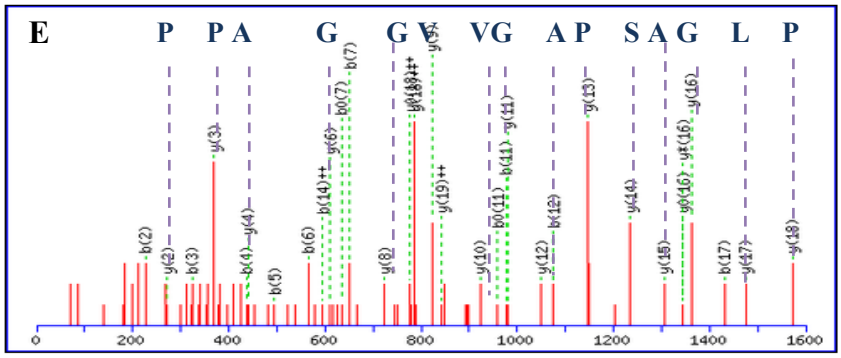
#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	129.0659	65.0366	112.0393	56.5233			Q							13
2	257.1244	129.0659	240.0979	120.5526			Q	1452.7230	726.8651	1435.6965	718.3519	1434.7124	717.8599	12
3	385.1830	193.0951	368.1565	184.5819			Q	*1324.6644*	662.8359	1307.6379	654.3226	1306.6539	653.8306	11
4	499.2259	250.1166	482.1994	241.6033			N	*1196.6058*	598.8066	1179.5793	590.2933	1178.5953	589.8013	10
5	586.2580	293.6326	569.2314	285.1193	568.2474	284.6273	S	*1082.5629*	541.7851	1065.5364	533.2718	1064.5524	532.7798	9
6	700.3009	350.6541	683.2743	342.1408	682.2903	341.6488	N	*995.5309*	498.2691	978.5043	489.7558	977.5203	489.2638	8
7	813.3850	407.1961	796.3584	398.6828	795.3744	398.1908	I	*881.4880*	441.2476	864.4614	432.7343	863.4774	432.2423	7
8	960.4534	480.7303	943.4268	472.2170	942.4428	471.7250	F	*768.4039*	384.7056	751.3774	376.1923	750.3933	375.7003	6
9	1107.5218	554.2645	1090.4952	545.7513	1089.5112	545.2592	F	*621.3355*	311.1714	604.3089	302.6581	603.3249	302.1661	5
10	1220.6058	610.8066	1203.5793	602.2933	1202.5953	601.8013	L	*474.2671*	237.6372	457.2405	229.1239	456.2565	228.6319	4
11	1291.6430	646.3251	1274.6164	637.8118	1273.6324	637.3198	A	*361.1830*	181.0951	344.1565	172.5819	343.1724	172.0899	3
12	1406.6699	703.8386	1389.6434	695.3253	1388.6593	694.8333	D	290.1459	145.5766	273.1193	137.0633	272.1353	136.5713	2
13							R	*175.1190*	88.0631	158.0924	79.5498			1

C) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide QQQNSNIFFALDR, found in the uORF of *ASNSD1* mRNA sequence.



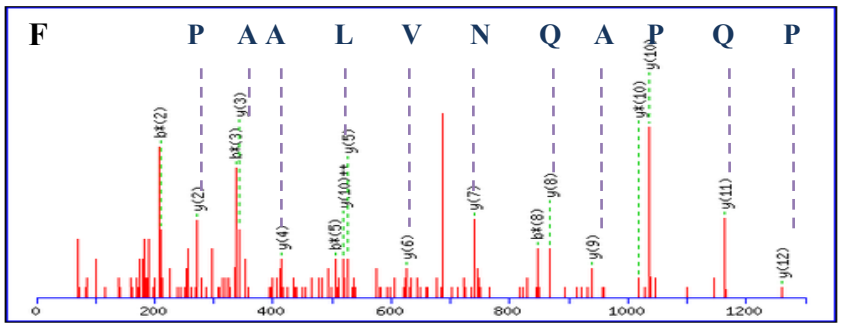
#	b	b^{++}	b*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	130.0499	65.5286			112.0393	56.5233	E							10
2	293.1132	147.0602			275.1026	138.0550	Y	1151.5579	576.2826	1134.5313	567.7693	1133.5473	567.2773	9
3	421.1718	211.0895	404.1452	202.5763	403.1612	202.0842	Q	*988.4946*	494.7509	971.4680	486.2376	970.4840	485.7456	8
4	550.2144	275.6108	533.1878	267.0975	532.2038	266.6055	E	*860.4360*	430.7216	843.4094	422.2084	842.4254	421.7163	7
5	663.2984	332.1529	646.2719	323.6396	645.2879	323.1476	I	*731.3934*	366.2003	714.3668	357.6871	713.3828	357.1951	6
6	792.3410	396.6742	775.3145	388.1609	774.3305	387.6689	E	*618.3093*	309.6583	601.2828	301.1450	600.2988	300.6530	5
7	906.3840	453.6956	889.3574	445.1823	888.3734	444.6903	N	*489.2667*	245.1370	472.2402	236.6237	471.2562	236.1317	4
8	1019.4680	510.2376	1002.4415	501.7244	1001.4575	501.2324	L	*375.2238*	188.1155	358.1973	179.6023	357.2132	179.1103	3
9	1134.4950	567.7511	1117.4684	559.2378	1116.4844	558.7458	D	*262.1397*	131.5735	245.1132	123.0602	244.1292	122.5682	2
10							K	*147.1128*	74.0600	130.0863	65.5468			1

D) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide EYQEIENLDK, found in the uORF of *ASNSD1* mRNA sequence.



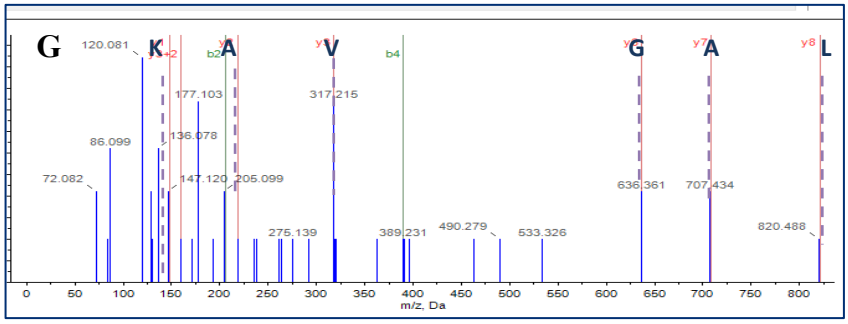
#	b	b^{++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	114.0913	57.5493			L							20
2	227.1754	114.0913			L	1685.9697	843.4885	1668.9432	834.9752	1667.9592	834.4832	19
3	324.2282	162.6177			P	*1572.8857*	786.9465	1555.8591	778.4332	1554.8751	777.9412	18
4	437.3122	219.1598			L	*1475.8329*	738.4201	1458.8063	729.9068	1457.8223	729.4148	17
5	494.3337	247.6705			G	*1362.7488*	681.8781	1345.7223	673.3648	1344.7383	672.8728	16
6	565.3708	283.1890			A	*1305.7274*	653.3673	1288.7008	644.8540	1287.7168	644.3620	15
7	652.4028	326.7051	634.3923	317.6998	S	*1234.6902*	617.8488	1217.6637	609.3355	1216.6797	608.8435	14
8	749.4556	375.2314	731.4450	366.2262	P	*1147.6582*	574.3327	1130.6317	565.8195			13
9	820.4927	410.7500	802.4822	401.7447	A	*1050.6055*	525.8064	1033.5789	517.2931			12
10	877.5142	439.2607	859.5036	430.2554	G	*979.5683*	490.2878	962.5418	481.7745			11
11	976.5826	488.7949	958.5720	479.7897	V	*922.5469*	461.7771	905.5203	453.2638			10
12	1075.6510	538.3291	1057.6404	529.3239	V	*823.4785*	412.2429	806.4519	403.7296			9
13	1132.6725	566.8399	1114.6619	557.8346	G	*724.4100*	362.7087	707.3835	354.1954			8
14	1189.6939	595.3506	1171.6834	586.3453	G	667.3886	334.1979	650.3620	325.6847			7
15	1246.7154	623.8613	1228.7048	614.8561	G	*610.3671*	305.6872	593.3406	297.1739			6
16	1359.7995	680.4034	1341.7889	671.3981	L	553.3457	277.1765	536.3191	268.6632			5
17	1430.8366	715.9219	1412.8260	706.9166	A	*440.2616*	220.6344	423.2350	212.1212			4
18	1527.8893	764.4483	1509.8788	755.4430	P	*369.2245*	185.1159	352.1979	176.6026			3
19	1624.9421	812.9747	1606.9315	803.9694	P	*272.1717*	136.5895	255.1452	128.0762			2
20					R	175.1190	88.0631	158.0924	79.5498			1

E) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide LLPLGASPAGVVGGGLAPPR, found in *TUBA3FP* ncRNA sequence



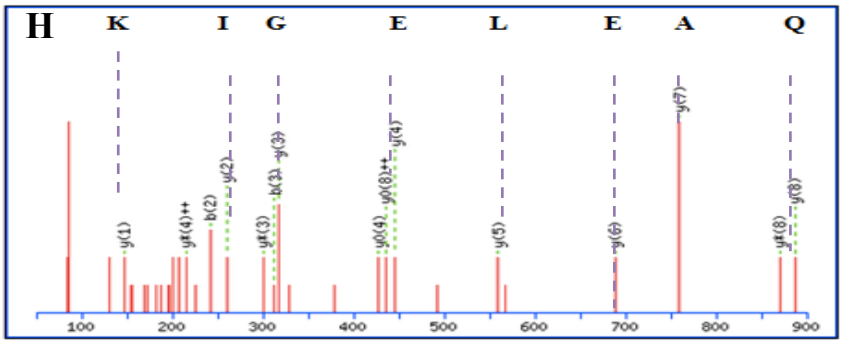
#	b	b^{++}	b*	b^{*++}	Seq.	y	y^{++}	y*	y^{*++}	#
1	129.0659	65.0366	112.0393	56.5233	Q					13
2	226.1186	113.5629	209.0921	105.0497	P	*1261.7011*	631.3542	1244.6746	622.8409	12
3	354.1772	177.5922	337.1506	169.0790	Q	*1164.6484*	582.8278	1147.6218	574.3146	11
4	451.2300	226.1186	434.2034	217.6053	P	*1036.5898*	518.7985	1019.5633	510.2853	10
5	522.2671	261.6372	505.2405	253.1239	A	*939.5370*	470.2722	922.5105	461.7589	9
6	650.3257	325.6665	633.2991	317.1532	Q	*868.4999*	434.7536	851.4734	426.2403	8
7	764.3686	382.6879	747.3420	374.1747	N	*740.4413*	370.7243	723.4148	362.2110	7
8	863.4370	432.2221	846.4104	423.7089	V	*626.3984*	313.7028	609.3719	305.1896	6
9	976.5211	488.7642	959.4945	480.2509	L	*527.3300*	264.1686	510.3035	255.6554	5
10	1047.5582	524.2827	1030.5316	515.7694	A	*414.2459*	207.6266	397.2194	199.1133	4
11	1118.5953	559.8013	1101.5687	551.2880	A	*343.2088*	172.1081	326.1823	163.5948	3
12	1215.6480	608.3277	1198.6215	599.8144	P	*272.1717*	136.5895	255.1452	128.0762	2
13					R	175.1190	88.0631	158.0924	79.5498	1

F) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide QPQPAQNVLAAPR, found in the uORF of *POLR2M* mRNA sequence.



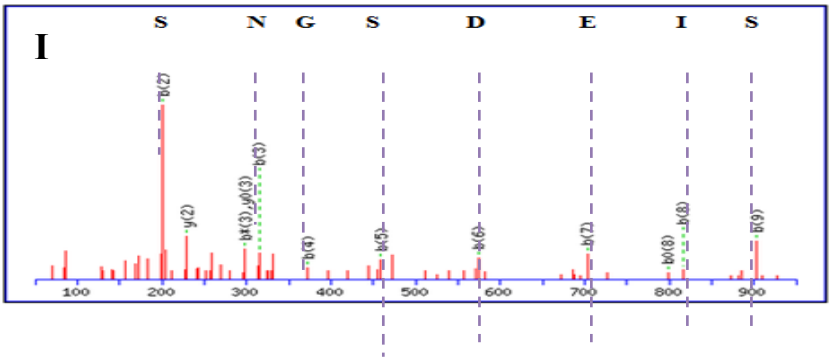
Residue	b	b+2	y	y+2
G	58.0287	29.5180	1024.5826	512.7949
F	**205.0972**	103.0522	967.5611	484.2842
L	318.1812	159.5942	**820.4927**	410.7500
A	**389.2183**	195.1128	**707.4087**	354.2080
G	446.2398	223.6235	**636.3715**	318.6894
Y	609.3031	305.1552	579.3501	290.1787
V	708.3715	354.6894	416.2867	208.6470
V	807.4400	404.2236	**317.2183**	**159.1128**
A	878.4771	439.7422	**218.1499**	109.5786
K	1006.5720	503.7897	**147.1128**	74.0600

G) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide GFLAGYVVAK, found in the oORF of *SLC35A4* mRNA sequence.



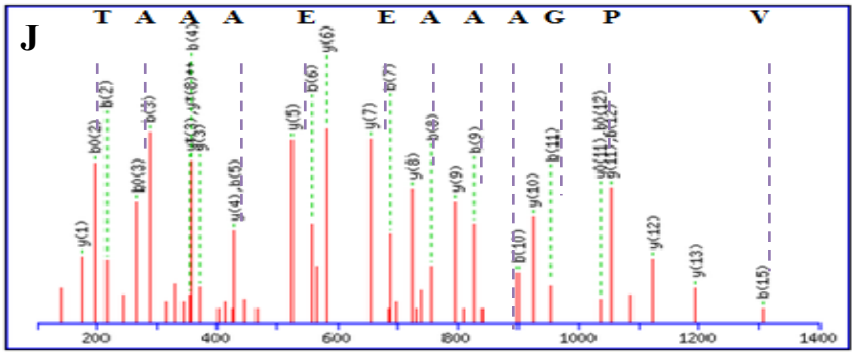
#	b	b^{++}	b*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	114.0913	57.5493					L							9
2	***242.1499***	121.5786	225.1234	113.0653			Q	***887.4833***	444.2453	870.4567	435.7320	869.4727	435.2400	8
3	***313.1870***	157.0972	296.1605	148.5839			A	***759.4247***	380.2160	742.3981	371.7027	741.4141	371.2107	7
4	442.2296	221.6185	425.2031	213.1052	424.2191	212.6132	E	***688.3876***	344.6974	671.3610	336.1842	670.3770	335.6921	6
5	555.3137	278.1605	538.2871	269.6472	537.3031	269.1552	L	***559.3450***	280.1761	542.3184	271.6629	541.3344	271.1709	5
6	684.3563	342.6818	667.3297	334.1685	666.3457	333.6765	E	***446.2609***	223.6341	429.2344	215.1208	428.2504	214.6288	4
7	741.3777	371.1925	724.3512	362.6792	723.3672	362.1872	G	***317.2183***	159.1128	300.1918	150.5995			3
8	854.4618	427.7345	837.4353	419.2213	836.4512	418.7293	I	***260.1969***	130.6021	243.1703	122.0888			2
9							K	***147.1128***	74.0600	130.0863	65.5468			1

H) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide LQAELEGIK, derived from the oORF of *NCAM2* mRNA sequence.



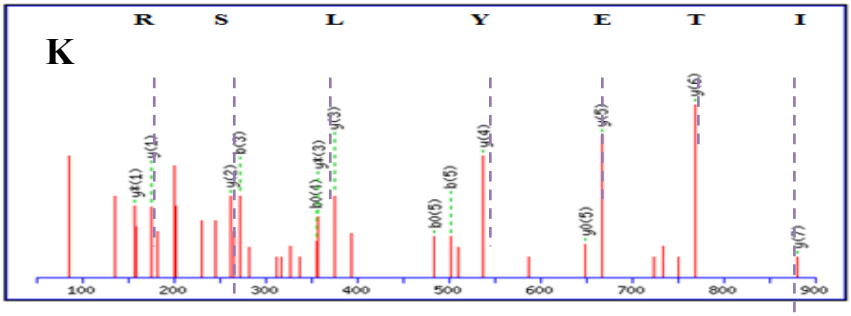
#	b	b^{++}	b*	b^{*++}	b^0	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^0	y^{0++}	#
1	114.0913	57.5493					I							11
2	***201.1234***	101.0653			183.1128	92.0600	S	1018.4687	509.7380	1001.4422	501.2247	1000.4582	500.7327	10
3	***315.1663***	158.0868	298.1397	149.5735	297.1557	149.0815	N	931.4367	466.2220	914.4102	457.7087	913.4262	457.2167	9
4	***372.1878***	186.5975	355.1612	178.0842	354.1772	177.5922	G	817.3938	409.2005			799.3832	400.1952	8
5	***459.2198***	230.1135	442.1932	221.6003	441.2092	221.1082	S	760.3723	380.6898			742.3618	371.6845	7
6	***574.2467***	287.6270	557.2202	279.1137	556.2362	278.6217	D	673.3403	337.1738			655.3297	328.1685	6
7	***703.2893***	352.1483	686.2628	343.6350	685.2788	343.1430	E	558.3134	279.6603			540.3028	270.6550	5
8	***816.3734***	408.6903	799.3468	400.1771	798.3628	399.6850	I	429.2708	215.1390			411.2602	206.1337	4
9	***903.4054***	452.2063	886.3789	443.6931	885.3948	443.2011	S	316.1867	158.5970			298.1761	149.5917	3
10	1016.4895	508.7484	999.4629	500.2351	998.4789	499.7431	L	***229.1547***	115.0810					2
11							P	116.0706	58.5389					1

I) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide ISNGSDEISLP, derived from *LINC00950* long ncRNA.



#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	114.0550	57.5311			A							16
2	*215.1026*	108.0550	197.0921	99.0497	T	1367.6914	684.3493	1350.6648	675.8360	1349.6808	675.3440	15
3	*286.1397*	143.5735	268.1292	134.5682	A	1266.6437	633.8255	1249.6171	625.3122	1248.6331	624.8202	14
4	*357.1769*	179.0921	339.1663	170.0868	A	*1195.6066*	598.3069	1178.5800	589.7937	1177.5960	589.3016	13
5	*428.2140*	214.6106	410.2034	205.6053	A	*1124.5695*	562.7884	1107.5429	554.2751	1106.5589	553.7831	12
6	*557.2566*	279.1319	539.2460	270.1266	E	*1053.5324*	527.2698	1036.5058	518.7565	1035.5218	518.2645	11
7	*686.2992*	343.6532	668.2886	334.6479	E	*924.4898*	462.7485	907.4632	454.2352	906.4792	453.7432	10
8	*757.3363*	379.1718	739.3257	370.1665	A	*795.4472*	398.2272	778.4206	389.7139			9
9	*828.3734*	414.6903	810.3628	405.6851	A	*724.4101*	362.7087	707.3835	354.1954			8
10	*899.4105*	450.2089	881.3999	441.2036	A	*653.3729*	327.1901	636.3464	318.6768			7
11	*956.4320*	478.7196	938.4214	469.7143	G	*582.3358*	291.6715	565.3093	283.1583			6
12	*1053.4847*	527.2460	1035.4742	518.2407	P	*525.3144*	263.1608	508.2878	254.6475			5
13	1110.5062	555.7567	1092.4956	546.7515	G	*428.2616*	214.6344	411.2350	206.1212			4
14	1207.5590	604.2831	1189.5484	595.2778	P	*371.2401*	186.1237	354.2136	177.6104			3
15	*1306.6274*	653.8173	1288.6168	644.8120	V	274.1874	137.5973	257.1608	129.0840			2
16					R	*175.1190*	88.0631	158.0924	79.5498			1

J) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide ATAAAEEAAAGPGPVR, found in the oORF of *TM9SF3* mRNA sequence.



#	b	b^{++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y^{*}	y^{*++}	y^{0}	y^{0++}	#
1	58.0287	29.5180			G							9
2	159.0764	80.0418	141.0659	71.0366	T	982.5204	491.7638	965.4938	483.2506	964.5098	482.7585	8
3	272.1605	136.5839	254.1499	127.5786	I	*881.4727*	441.2400	864.4462	432.7267	863.4621	432.2347	7
4	373.2082	187.1077	355.1976	178.1024	T	*768.3886*	384.6980	751.3621	376.1847	750.3781	375.6927	6
5	502.2508	251.6290	484.2402	242.6237	E	*667.3410*	334.1741	650.3144	325.6608	649.3304	325.1688	5
6	665.3141	333.1607	647.3035	324.1554	Y	*538.2984*	269.6528	521.2718	261.1396	520.2878	260.6475	4
7	778.3981	389.7027	760.3876	380.6974	L	*375.2350*	188.1212	358.2085	179.6079	357.2245	179.1159	3
8	865.4302	433.2187	847.4196	424.2134	S	*262.1510*	131.5791	245.1244	123.0659	244.1404	122.5738	2
9					R	*175.1190*	88.0631	158.0924	79.5498			1

K) MS/MS spectrum of the sPEP identified by proteomic procedure matching the peptide GTITEYLSR, found in the oORF of *PTPN21* mRNA sequence.

Figure 3.12. The MS/MS raw data & tables of sequence matches from fragment ions of the identified sPEPs.

A) MS/MS spectrum of the sPEP, IVVDELSNLK, found in the uORF of *ASNSD1* mRNA sequence. B) MS/MS spectrum of the sPEP, NILDELK, found in the uORF of *ASNSD1* mRNA sequence. C) MS/MS spectrum of the sPEP, QQQNSNIFFALDR, found in the uORF of *ASNSD1* mRNA sequence. D) MS/MS spectrum of the sPEP, EYQEIENLDK, found in the uORF of *ASNSD1* mRNA sequence. E) MS/MS spectrum of the sPEP, LLPLGASPAGVVGGGLAPPR, found in *TUBA3FP* ncRNA sequence. F) MS/MS spectrum of the sPEP, QPQPAQNVLAAPR, found in the uORF of *POLR2M* mRNA sequence. G) MS/MS spectrum of the sPEPE, GFLAGYVVAK, found in the oORF of *SLC35A4* mRNA sequence. H) MS/MS spectrum of the sPEP, LQAELEGIK, derived from the oORF of *NCAM2* mRNA sequence. I) MS/MS spectrum of the sPEP, ISNGSDEISLP, derived from *LINC00950* long ncRNA. J) MS/MS spectrum of the sPEP, ATAAAEEAAAGPGPVR, found in the oORF of *TM9SF3* mRNA sequence. K) MS/MS spectrum of the sPEP, GTITEYLSR, found in the oORF of *PTPN21* mRNA sequence.

Discussion

Experiments had been performed repeatedly with minor changes in the protocol, including cell culturing and protein enrichment strategies to reconfirm those sPEPs found in previous experiments and also to obtain more data for sPEP identification. In order to prevent protein degradation during cell lysis, 8-hour 50 μM PI treatment was applied on HEK293 cells while 4-hour 50 μM PI was treated in HeLa cells before lysis. Results showed that the amount of protein products obtained was doubled in both HeLa and HEK293 cells with 50 μM PI treatment (Figure 3.4 & 3.5). From the results of protein identification in MASCOT searches, no significant increase in sPEP detection in both cell lines with PI treatment (Figure 3.6). In the results of MS/MS analysis, the same sPEP (peptide sequence: ISNGSDEISLP, derived from *LINC00950* long ncRNA) appeared from the samples with and without PI treatment. Results indicated that PI treatment in cells could benefit in protein yield and it also supported that PI bortezomib does interfere with protein degradation (Gelman et al., 2013). However, in my result, the efficiency in sPEP identification after the use of PI bortezomib had no noticeable increase in the amount of protein yield. This may be due to the loss

of protein during protein separation processes, or the availability in the detection of peptides in LC-MS/MS analysis.

Among the 11 identified sPEPs in my result, three of these have been confirmed as novel sPEPs. The role of these novel sPEPs remains unknown. Therefore, further characterisation of these sPEPs would be beneficial to explore their potential function.

Chapter 4

Results from bioinformatic approaches

Introduction

Bioinformatic analyses provide critical information for genes or proteins of interest by predicting the function of actual gene products. There are various bioinformatic tools for data analysis. For example, in molecular biology, tools which are designed to predict the secondary structure of a protein provides essential intermediate information on the way to predicting the full 3-D structure of a protein, and hence to predicting its function. Cross-species conservation of sORFs can reveal those that encode potential functionally important peptides, since high levels of sequence identity between sORF orthologues are an indication that their encoded uPEP has been maintained during evolution (Crowe et al., 2006).

Analysis of sORFs for cross-species conservation

sORFs identified in both bioinformatic and proteomic studies were analysed for conservations in other species using both of Blast NCBI and uPEPperoni online tools, which detect conserved uORFs in eukaryotic transcripts. A NCBI BLAST search of both nucleotide and amino acid sequences of the sPEP from *LINC00950* ncRNA revealed conservation between human, mouse, rat, and zebra-fish (Appendix 3a). The sPEP from the uORF of *TM9SF3* was found to be conserved in human, mouse, bovine, and macaque (Appendix 3b), while the other sPEP from oORF of *PTPN21* showed conservations between human, mouse and rat (Appendix 3c). The high degree of conservation suggests that these sPEPs could be functional.

Bioinformatic analysis of the characterisation of the novel sORFs

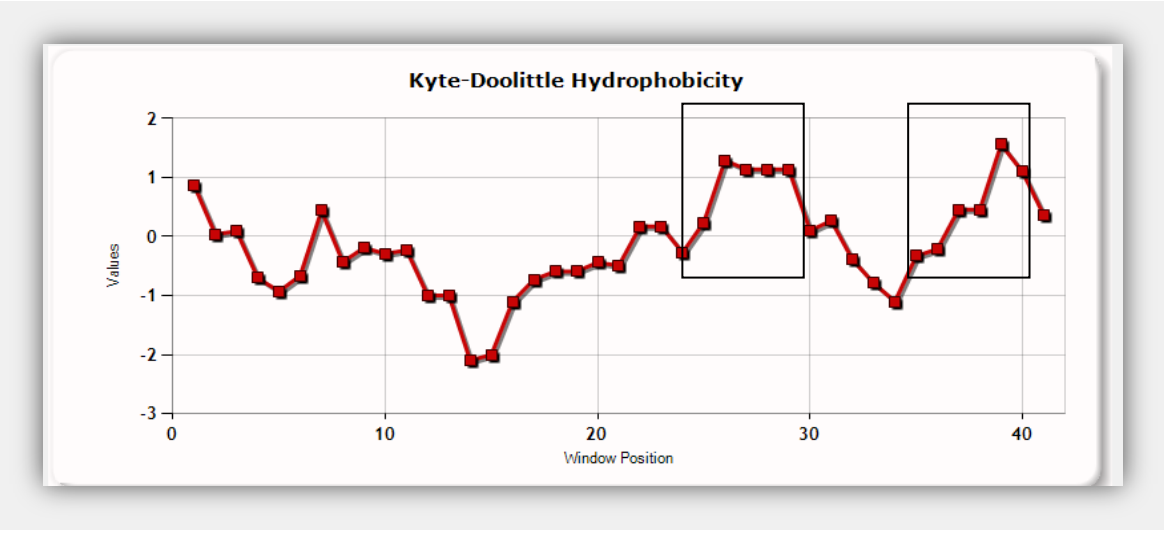
Characterisation of the identified sORFs was obtained using a web-based peptide/protein analysis tool to look at properties of the sPEP such as protein hydrophobicity, transmembrane region predictions, and protein flexibility (Figure 4.1 - 4.3). Characterisation supports the potential that

these novel sPEPs can be bioactive and functional in cells due to particular features that are representative to be functional peptides such as having amphiphilicity helices, transmembrane regions (Appendix 4). Proteins play a significant role in biological processes by various folding possibilities. The prediction of a protein's secondary structure is an essential intermediate step on the way to predicting the full 3-D structure of a protein by packing secondary structure elements into globular domains (Chen and Lonardi, 2009). Therefore, secondary structure predictions of the 3 novel sPEPs have been analysed by bioinformatic tools. From the results, the low prediction confidence for the secondary structure of the sPEPs obtained through Phyer^2 could be due to the short peptide sequence used for query searches (Appendix 5). Protein localisation prediction of these sPEPs was analysed using protein structure prediction on the web: The PredictProtein server Rost, LocTree2, which predicts the subcellular localisation of all domains (Appendix 6). The novel sPEP from *LINC00950* ncRNA is expected to be secreted. The sPEP from uORF of *TM9SF3* is predicted to localise to the cytoplasm while the sPEP from oORF of *PTPN21* is expected to be secreted.

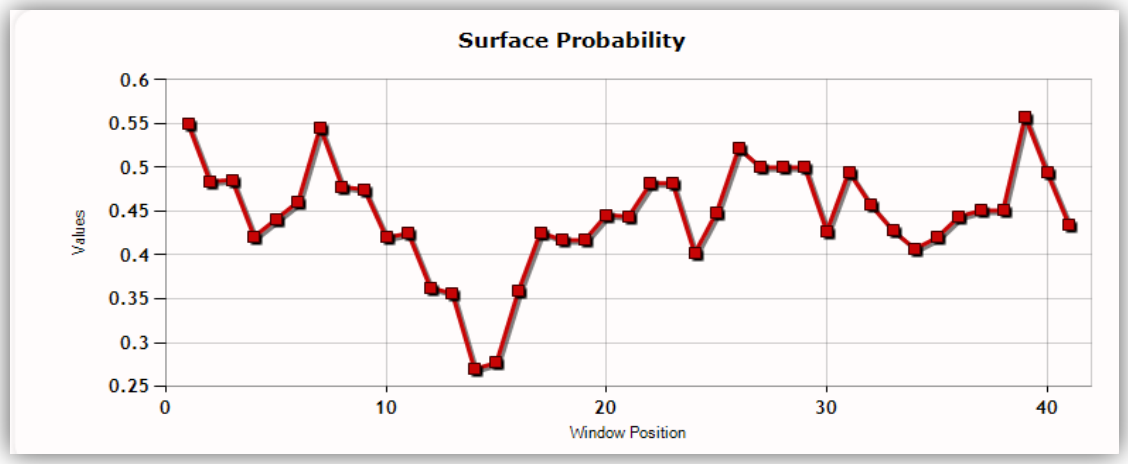
Gene expression in cells plays an important in the investigation of the synthesis of a functional product. Protein expression of the novel sPEPs identified in this project was analysed using a web-based gene and protein function analysis tool: BioGPF. From the results, *LINC00950* ncRNA is found mainly in the superior cervical ganglion (Appendix 7a), which gives the protein expression prediction of the novel sPEP from this ncRNA. The sPEP from uORF of *TM9SF3* (Appendix 7b) is predicted to have high expression in the colon while the sPEP from oORF of *PTPN21* is predicted to be associated in the functions in cardiac myocytes, lymphoma, and liver (Appendix 7c). Protein motifs are small elements that are often used to predict protein function. Motifs are conserved among different proteins and may have structural or functional roles. Motif prediction for the three novel sPEPs were performed using MyHit, a database website, which analyses a protein sequence

for the known motifs through the program PROSITE. The *LINC00950* sPEP sequence analysed has a potential N-glycosylation site at amino acid 29 (Appendix 8a). For *TM9SF3* sPEP, it shows a strong match in the PROSITE database to have a potential alanine-rich region profile at amino acid 9 (Appendix 8b). The result for this sPEP also shows several weak matches to have amidation sites, N-myristoylation sites, and protein kinase C phosphorylation sites. The *PTPN21* sPEP sequence analysed shows a potential N-glycosylation site at amino acid 28, two casein kinase II phosphorylation sites at amino acid 9 & 30, two N-myristoylation sites at amino acid 21 & 37, and a and protein kinase C phosphorylation at amino acid 34 (Appendix 8c). With these potential chemical groups in sPEPs, it may imply that these sPEPs could be functional and associated in biological mechanisms, such as altering protein folding and stability; thus, regulating of protein function.

A



B



C

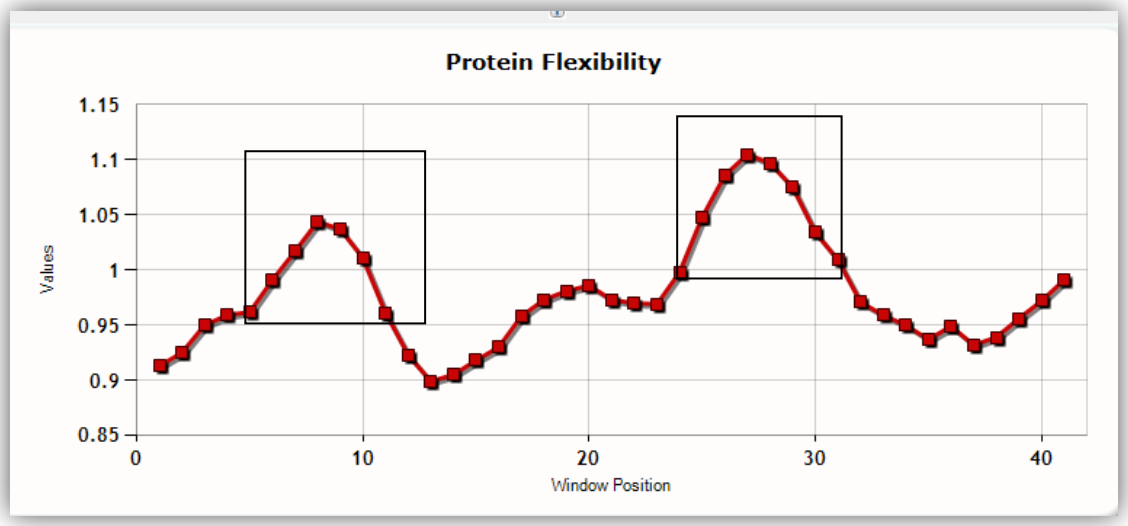
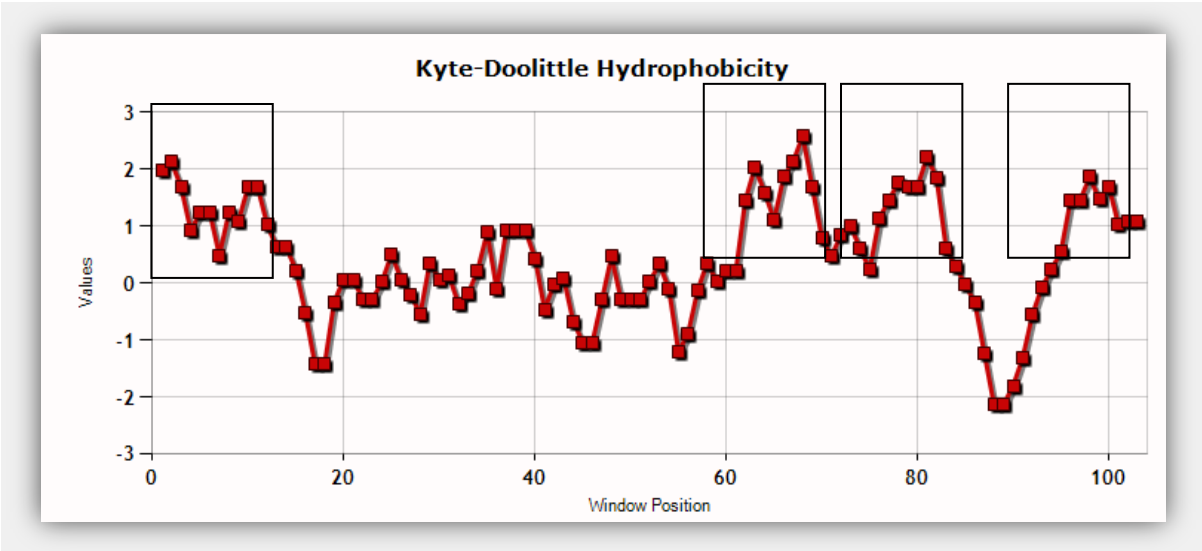


Figure 4.1. Predictions of hydrophobicity, hydrophilicity, and flexibility of the sPEP from *LINC00950* ncRNA

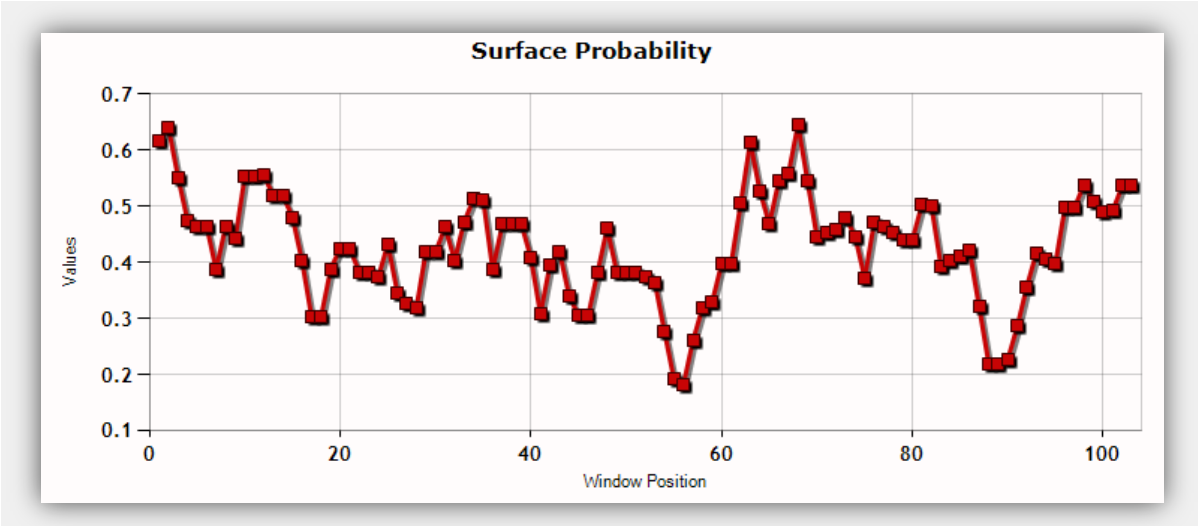
A) The graph is used to find clusters of hydrophobic amino acids or find potential antigenic sites of globular proteins. The transmembrane regions or antigenic regions of the peptide are seen as distinct peaks in the graph (labeled in black circles). B) This graph shows the prediction of the surface and flexibility properties regions of the sPEP from *LINC00950* ncRNA. The sum of six fractional probabilities of the amino acids in the window are taken at once and divided by six to yield a running average of the fractional surface probability along the length of the protein. A value of 1.0 at any point (which will never occur) would mean that the hexapeptide centered about that point is definitely exposed at the surface of the protein and a value of 0.0 (which also will never

occur) means that the hexapeptide is definitely buried in the interior of the peptide. For this peptide, the hexapeptide is in the intermediate area of the protein. C) The average flexibility of a protein is 1.0. Regions with values greater than 1.0 are predicted to be more flexible and values below 1.0 indicate regions predicted to be less flexible as compared to average flexibility of the protein. There are two distinct regions (labeled in black circles) of this peptide that are predicted to be have average flexibility.

A



B



C

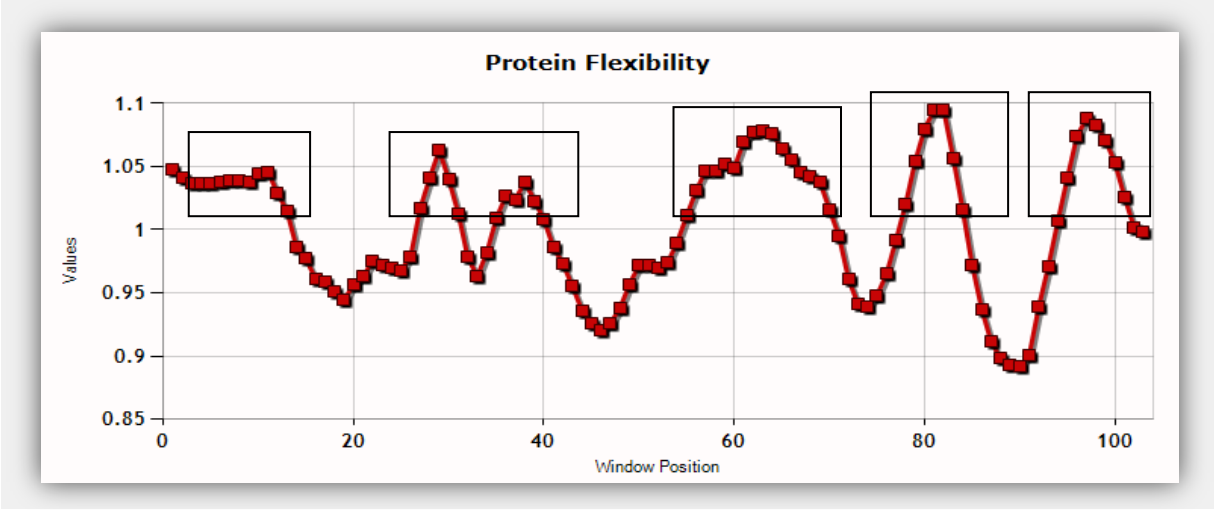
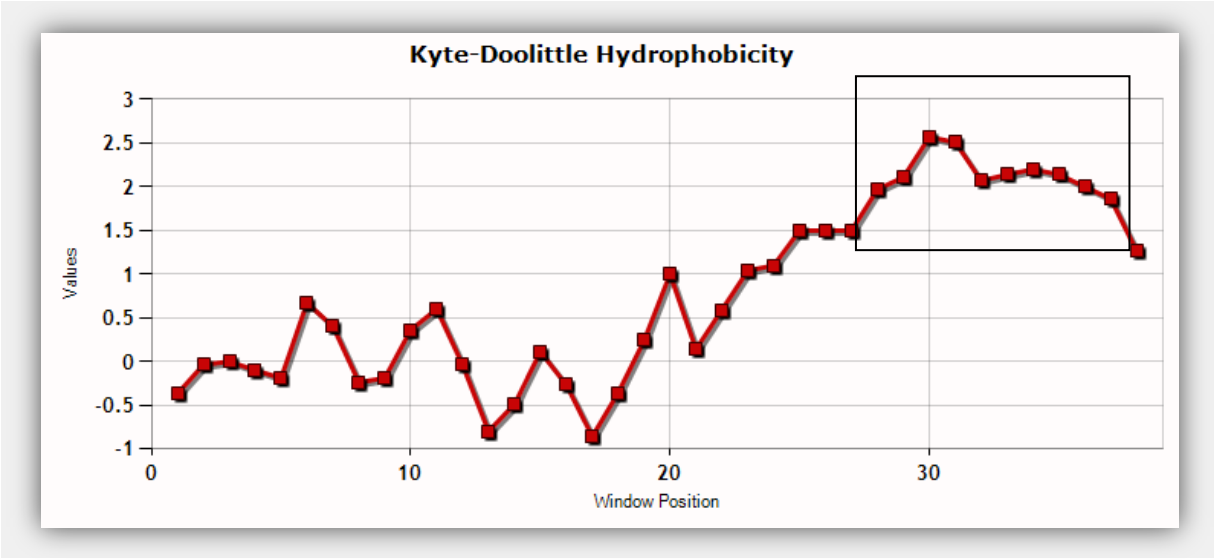


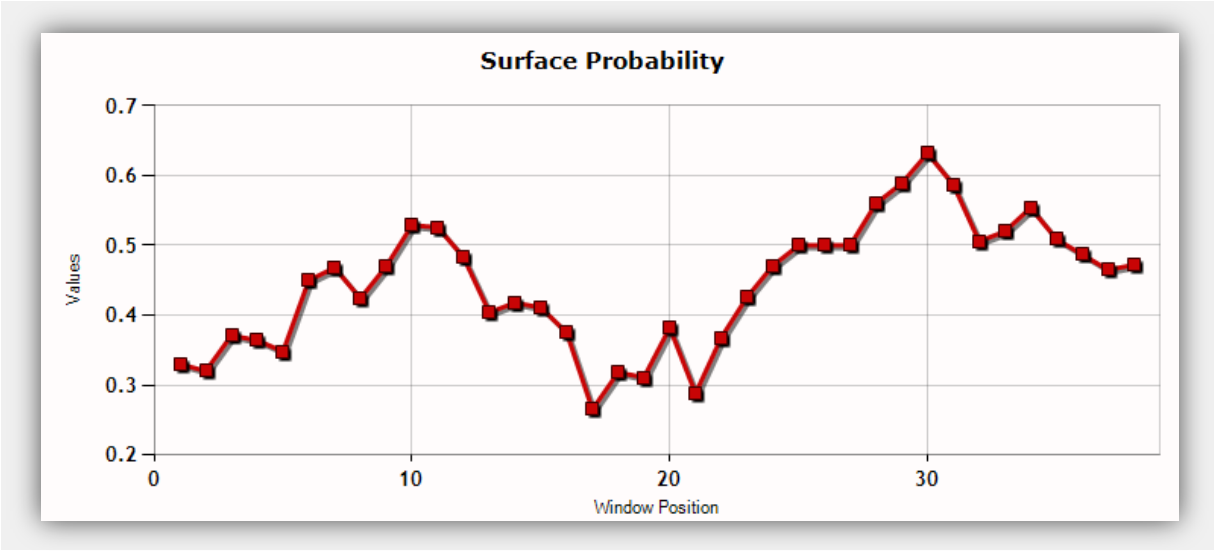
Figure 4.2. Predictions of hydrophobicity, hydrophilicity, and flexibility of the sPEP from the uORF of *TM9SF3*

A) The graph is used to find clusters of hydrophobic amino acids or find potential antigenic sites of globular proteins. The transmembrane regions or antigenic regions of the peptide are seen as distinct peaks in the graph (labeled in black circles). B) This graph shows the prediction of the surface and flexibility properties regions of the sPEP from the uORF of *TM9SF3*. The sum of six fractional probabilities of the amino acids in the window are taken at once and divided by six to yield a running average of the fractional surface probability along the length of the protein. A value of 1.0 at any point (which will never occur) would mean that the hexapeptide centered about that point is definitely exposed at the surface of the protein and a value of 0.0 (which also will never occur) means that the hexapeptide is definitely buried in the interior of the peptide. For this peptide, the hexapeptide is in the intermediate area of the protein. C) The average flexibility of a protein is 1.0. Regions with values greater than 1.0 are predicted to be more flexible and values below 1.0 indicate regions predicted to be less flexible as compared to average flexibility of the protein. There are five distinct regions (labeled in black circles) of this peptide that are predicted to be have average flexibility.

A


Kyte-Doolittle Hydrophobicity
Values
3
2.5
2
1.5
1
0.5
0
-0.5
-1
0
10
20
30
Window Position

B


Surface Probability
Values
0.7
0.6
0.5
0.4
0.3
0.2
0
10
20
30
Window Position

C

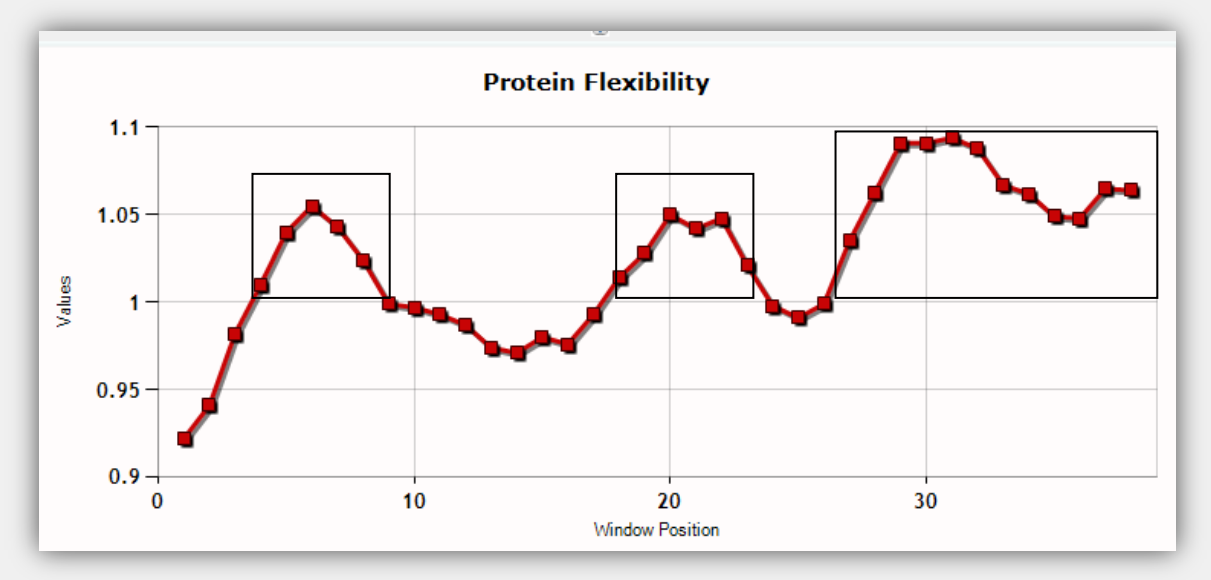


Figure 4.3. Predictions of hydrophobicity, hydrophilicity, and flexibility of the sPEP from the oORF of *PTPN21*

A) The graph is used to find clusters of hydrophobic amino acids or find potential antigenic sites of globular proteins. The transmembrane regions or antigenic regions of the peptide are seen as distinct peaks in the graph (labeled in black circle). B) This graph shows the prediction of the surface and flexibility properties regions of the sPEP from the oORF of *PTPN21*. The sum of six fractional probabilities of the amino acids in the window are taken at once and divided by six to yield a running average of the fractional surface probability along the length of the protein. A value of 1.0 at any point (which will never occur) would mean that the hexapeptide centered about that point is definitely exposed at the surface of the protein and a value of 0.0 (which also will never occur) means that the hexapeptide is definitely buried in the interior of the peptide. For this peptide, the hexapeptide is in the intermediate area of the protein. C) The average flexibility of a protein is 1.0. Regions with values greater than 1.0 are predicted to be more flexible and values below 1.0 indicate regions predicted to be less flexible as compared to average flexibility of the protein. There are three distinct regions (labeled in black circles) of this peptide that are predicted to be have average flexibility.

Analysis of CGG-repeat uORFs in neural transcripts

Tandem repeats are common features of the genomes of prokaryotes and eukaryotes. They can be found in intergenic regions and also have been found in both non-coding and coding regions of different gene transcripts (Subirana and Messeguer, 2008). A number of human neurological disorders have been shown in association with nucleotide repeat expansions (Orr and Zoghbi, 2007). These dynamic mutations cause diseases by protein gain-of-function, protein loss-of-function, or RNA gain-of-function mechanisms and may occur in promoters, coding regions, introns and the 5'leader and 3'trailer regions of mRNAs (McMurray, 2010). For example, expansion of a CGG nucleotide repeat (55-200) in the 5' UTR or promoter regions of the human *FMR1* gene have been reported in association with Fragile X-associated Tremor Ataxia Syndrome (FXTAS), an inherited mental retardation (Penagarikano et al., 2007).

To explore traces of conservation between human CGG-repeat uORFs in neural transcripts with different species, uORFs (a neural uORF list from John Carson, Connecticut) containing CGG-repeat elements were analysed using uPEPperoni online search engine, which detects conserved uORFs in eukaryotic transcripts. There are 23 hits of conserved uPEPs in other species obtained in total 226 human neural uORFs from the list (Appendix 9). Examples from uPEPperoni were shown in Figure 4.4. uPEPperoni also provides a Ka/Ks ratio to discover the evidence of purifying, positive or neutral selection. In this tool, sequences can be required to be of similar length or located in the same position relative to a previously identified ORFs. Further analyses of these sPEPs are required for characterisation.

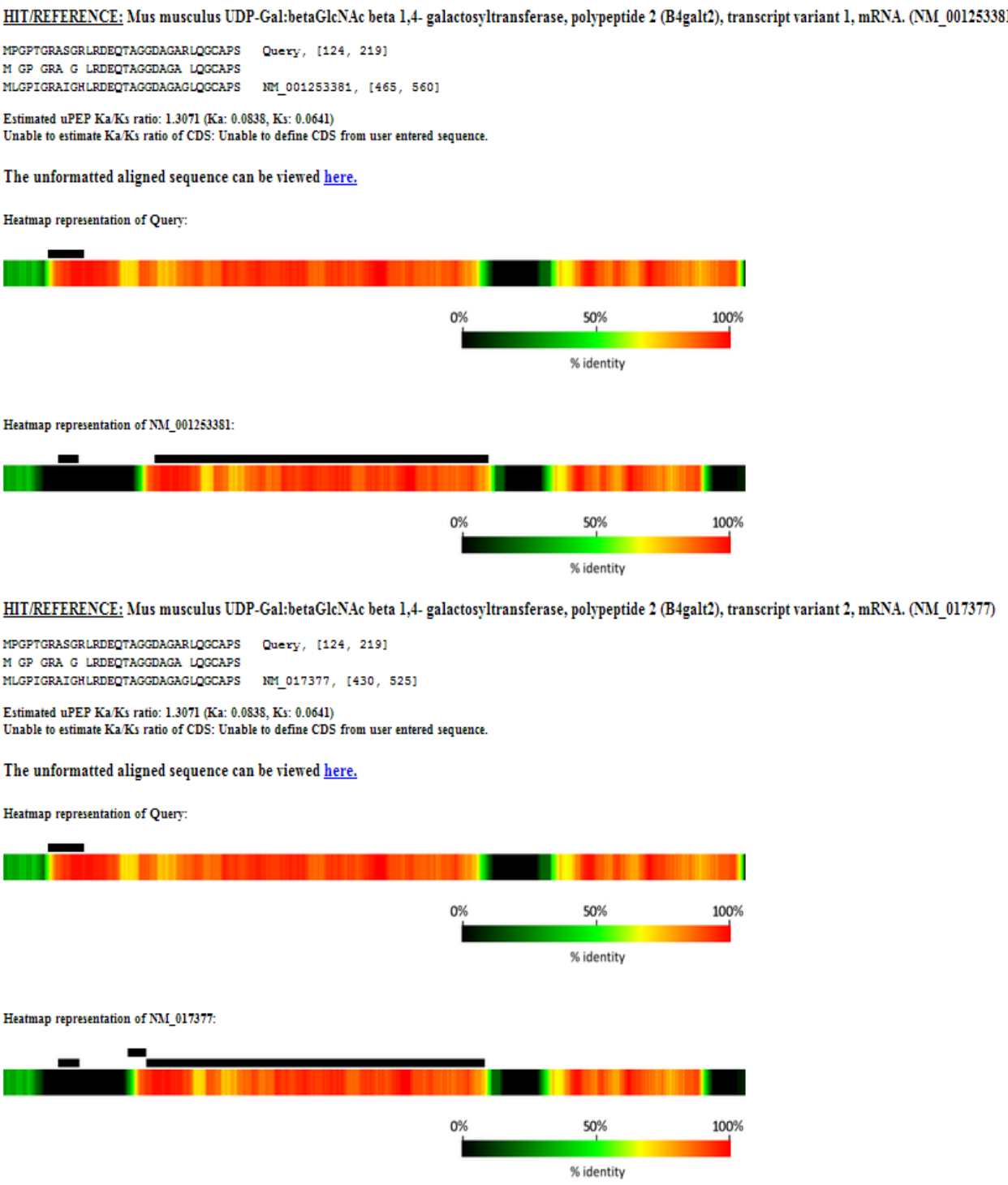


Figure 4.4. Example output showing the heatmaps produced by querying the mRNA sequence of the Homo sapiens *B4GALT2* transcript (NM_030587) against Mus musculus *B4GALT2*transcript variant 1 (NM_001253381) and transcript variant 2 (NM_017377) analysed from uPEPperoni online search engine

The black bars above the heatmap indicate the ORFs on the transcript. The output lists the most

conserved uPEPs first. The heatmap generated by the query sequence is shown first, followed by the reciprocal heatmap generated using the reference sequence (mouse *B4GALT2* transcript). The unformatted aligned sequence can be viewed using a hyperlink shown above the heatmap. uPEPperoni also provides Ka/Ks ratio for potential protein coding regions prediction (Skarshewski et al., 2014).

Analysis of sPEPs with four or more Cys residues

Cysteine is often found in the functional sites in proteins, participating in catalytic, regulatory, cofactor-binding, structure-stabilizing functions (Marino and Gladyshev, 2012, Tiessen et al., 2012). Cys mutations have been reported in association with the connections between human genetic diseases and evolution (Wu et al., 2007). To explore traces of bioactive sPEPs, sPEPs identified in both bioinformtic and proteomic studies were analysed for the presence of four or more Cys residues in sequence (Table 4.1). Five Cys residues were found to be prevalent in these sPEPs, followed by four and six Cys residues. Only one sPEP had 12 Cys residues (Figure 4.5). There are several methods that can be used to examine Cys reactivity, such as pKa measurements by identifying Cys residues with heightened nucleophilicity. However, this method may not be suitable for a proteomic scale because it requires purified protein and detailed kinetic and mutagenic experiments (Quantitative reactivity profiling predicts functional cysteines in proteomes). Alternative methods have been reported to predict redox-active cysteine computationally by identifying Cys with specific modifications (Sethuraman et al., 2004). Based on these methods, examination of those sPEPs containing four or more Cys is worthy for further investigation.

Discussion

Cross-species conservation of sORFs can reveal those that encode potential functionally important peptides, since high levels of sequence identity between sORF orthologues are an indication that their encoded uPEP has been maintained during evolution (Crowe et al., 2006). uPEPperoni also provides a Ka/Ks ratio to discover the evidence of purifying, positive or neutral

selection(Skarshewski et al., 2014). In this tool, sequences can be required to be of similar length or located in the same position relative to a previously identified ORFs. sORFs that lack cross-species conservation have been reported to be more likely random sequences without encoding for functional peptides (Andrews and Rothnagel, 2014). Therefore, it is important to obtain evidence of evolutionary conservation as a predictor of function. However, non-conserved sORFs are still worthy to be retained for further analyses since species-specific sPEPs may also be biologically relevant.

Although sPEPs can be identified bioinformatically, their function in cell biology needs more investigation. In order to understand the functionality of the identified sPEPs, characterisation through cellular and bioinformatic tools is essential. Computational biology is often used in predicting structures, membrane organisation or localisation signal of sPEPs. Complementary DNA of these sPEP sequences could be isolated and sequenced to observe their expression at both mRNA and protein levels. In particular, subcellular localisation of these sPEPs could be investigated through immunostaining experiments and using specific antibodies or fluorescent tags. Western blot and enhanced MS studies could be also implemented to explore the posttranslational processing of the identified sPEPs.

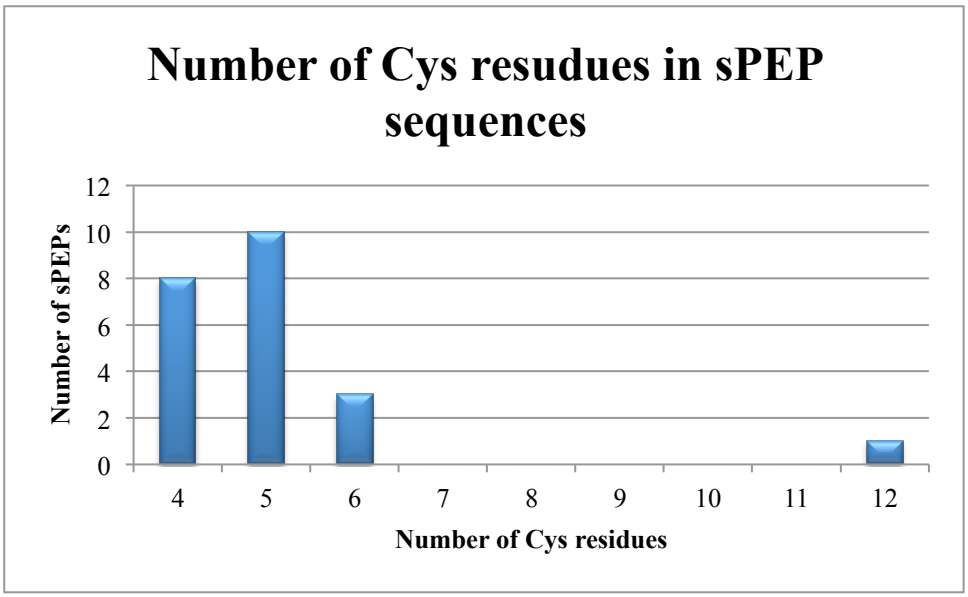


Figure 4.5. The prevalence of Cys residues contained in identified sPEPs

There are four Cys residues found in 8 sPEPs of the identified sPEPs, five Cys residues found in eight sPEPs, and six Cys residues in 3 sPEPs. Only one sPEP has 12 Cys residues.

NCBI Accession number	Chr	Strand	Coordinate gene name	Coordinate start-stop codon	uORF start-end	Peptide sequence
NM_007106_1	13	-	Homo sapiens ubiquitin-like 3 (UBL3), mRNA	29849867-29849805	817-882	MSVCHSARSTWRGR SWGCCCC
NM_015457_4	11	+	Homo sapiens zinc finger, DHHC-type containing 5 (ZDHHC5), mRNA	57672718-57672828	884-997	MSCILICLTVHVFHL QPFACVQPTVCLQFL NCTSCVS
NM_016192_2	2	-	Homo sapiens transmembrane protein with EGF-like and two follistatin-like domains 2 (TMEFF2), mRNA	192814004-192814090	236-322	MRGFGCCFPAGCHC HRRRLCCRRPRDAQ
NM_016513_1	6	-	Homo sapiens intestinal cell (MAK-like) kinase (ICK), transcript variant 2, mRNA	53041385-53041269	342-461	MYLGDSHVLLNTLC WCCHRKIWLHSGED CYHCRTEPLRP
NM_020665_1	X	-	Homo sapiens transmembrane protein 27 (TMEM27), mRNA	15664849-15664751	183-284	MAKADLSAAWIFFFS LCLVFSTLKECCGCS FFW

NM_024735_1	16	-	Homo sapiens F-box protein 31 (FBXO31), transcript variant 1, mRNA	87360363-87347185	393-530	MVFAKTCGSWRSQACLVGTSMRSCFTDIDTFWDCGSQISGHTEDC
NM_138447_1	16	-	Homo sapiens zinc finger protein 689 (ZNF689), transcript variant 1, mRNA	30610351-30610214	36-176	MALRSIKSIAGSCLCSRQRRCGSSAAIFPEGIFRCLSPKFGQEFPE
NM_178011_4	10	+	Homo sapiens leucine rich repeat transmembrane neuronal 3 (LRRTM3), mRNA	66926327-66926536	294-506	MKILLPRKILMFCCECGVGIYLFLECSAWLAKNNVPKSVHLPRGPIFLPGCQRALTHYSAADRGCHATGP
NM_031936.4	1	+	Homo sapiens G protein-coupled receptor 61 (GPR61), mRNA	109542971-109543336	632-1000	MGDGPVTGGTLGALFRPHGVLTHPPVIRELFHFGEGPSNPRSLYCQWGPGGGATGCCFGICGPLLHAPAGLDCCGWQCRCDGRDRQDACPPKICLRLPPLPGGPAGCPDPHAPGHALQLCPL
NM_015282.2	2	-	Homo sapiens cytoplasmic linker associated protein 1 (CLASP1), transcript variant 1, mRNA	121606051-121605929	235-360	MVHCCHSPDCIFETQALSNLQRTKRQPPRYVCWEGVIVTAL
NM_080670.2	5	+	Homo sapiens solute carrier family 35, member A4 (SLC35A4), mRNA	140566788-140566949	347-508	MSSSAFRWRHFIWIPQLSSICYLQLSCHLHPCLPSCRLWTVVPQPAPWIPSSPS
NM_002285.2	2	-	Homo sapiens AF4/FMR2 family, member 3 (AFF3), transcript variant 1, mRNA	99551326-99551198	4064-4193	MCVCMCGYIRSRLCMCVR
NM_019886.3	X	+	Homo sapiens carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 7 (CHST7), mRNA	46433686-46434177	494-986	ARCATCCVRSSAATSPCCGCTR

NM_182898.2	7	+	Homo sapiens cAMP responsive element binding protein 5 (CREB5), transcript variant 1, mRNA	28860728-28860865	3748-3886	LDLLMCVCVCVCVCVFMGFK
NM_014780.4	6	-	Homo sapiens cullin 7 (CUL7), transcript variant 2, mRNA	43018833-43017971	1408-1702	MLCMCGTHCSRGCECGCWMIMRR
NM_015089.2	6	+	Homo sapiens cullin 9 (CUL9), mRNA	43154066-43155076	1198-1555	MENMCSRHSSQGCECGCWMIMR
NM_004265.3	11	+	Homo sapiens fatty acid desaturase 2 (FADS2), transcript variant 1, mRNA	61865736-61866005	1531-1801	ARGMMGFCSEGCPRGWCMHCSR
NM_152312.3	11	+	Homo sapiens glycosyltransferase-like 1B (GYLTL1B), mRNA	45924226-45924588	551-686	MDGACCPCQLLSCR
NM_144612.6	18	-	Homo sapiens lipoxygenase homology domains 1 (LOXHD1), transcript variant 1, mRNA	46547017-46546874	2704-2848	MMASCPGSCCQWMSPMCCHRAR;MMASCPGSCCQWMSPMCCHR
NM_003791.2	16	-	Homo sapiens membrane-bound transcription factor peptidase, site 1 (MBTPS1), mRNA	84066566-84065717	2777-2906	MSCCLCGTWGSAMACMKGSSPWPTMTCIMR
NM_004959.4	9	-	Homo sapiens nuclear receptor subfamily 5, group A, member 1 (NR5A1), mRNA	124500613-124500077	533-1070	TATQSPSLEGPTCLSSSCSCCSWSRMR
NM_020151.3	2	-	Homo sapiens StAR-related lipid transfer (START) domain containing 7 (STARD7), mRNA	96185102-96184851	3141-3396	MWMDVCIRERENMCVCVCVCER

Table 4.1. sPEPs identified in both bioinformtic and proteomic studies with the presence of four or more Cys residues in sequences

sPEPs that contain four or more cystein residues were collected and arranged into a table listing their coordinate gene sequence (NCBI accession number), chromosomal location, DNA strand

orientation, the location of the start/stop codon of their coordinate sORF in chromosome, the location of the start/stop codon of their coordinate sORF, and the sPEP sequence identified in both bioinformatic and proteomic studies.

Chapter 5

General Discussion

Introduction

The aim of this project was to identify novel peptides that are translated from sORFs. We hypothesize that a subset of sORFs encode for functional sPEPs that are expressed and contribute to proteome complexity. Recent proteomic studies of endogenous proteins expressed in human cell lines have led to the discovery of novel sPEPs (Slavoff et al., 2013, Vanderperre et al., 2013). sPEPs are presumed to be small and of low abundance. In addition, the dynamic range and the complexity of the cellular proteome, such as post-translational modifications create challenges when sampling low abundant and small proteins (Nielsen et al., 2006). Therefore, it is necessary to develop effective, comprehensive, low-molecular-weight protein extraction methodologies to analyse endogenous proteins in a relatively complex and dynamic range of human cells.

Mass spectrometry-based methodologies in peptide enrichment approaches have been proven to be an ideal analytical method for mapping of peptides (Zarei et al., 2011). However, due to sample complexity, fractionation of peptides prior to MS analysis is a critical step. Two different peptide fractionation strategies were evaluated in this project. Although ERLIC and SCX approaches were demonstrated to have the best performance for detecting low abundance proteins in my experiments, it is still difficult to comment whether sPEPs can be better extracted by one particular method. Overall, both ERLIC and SCX approaches resulted more protein products from the same amount of starting material than that from SDS-PAGE gel approach. However, the MS/MS results did not show more sPEP identification through ERLIC and SCX approaches than that through SDS-PAGE gel approach. In addition, the same sPEPs were found multiple times regardless of method.

ERLIC has been shown to have better result in the separation of multi-phosphorylated peptides, while SCX is suited for the fractionation of mono-phosphorylated peptides (Zarei et al., 2012). A combination of ERLIC and SCX approach would increase the coverage of proteome analysis. As

the knowledge of endogenous sPEP expression is limited (structure, size, abundance etc.), combination of these peptide enrichment methods may be more favorable for sPEP detection in the future.

Validation of protein extraction methods

Molecular Weight Cut-Off (MWCO) + ERLIC (or SCX) approach

ERLIC has been reported to have 36% more protein identifications than via SCX fractionation (Sze et al., 2010). Moreover, over 120% more highly hydrophobic and basic peptides were identified by ERLIC than SCX (Sze et al., 2010). The average amount of protein products yield from ERLIC and SCX was calculated for comparison (Figure 3.8). From the result, protein identification via ERLIC fractionation was ~50% higher than via SCX, indicating that ERLIC operates higher sensitivity in protein separation and identification. The total number of proteins identified in each experiment varied quite a lot which may be due to the differences in the protocols, including cell culturing (PI treatment) and peptide separation and extraction strategies. From the results, the number of protein products obtained via ERLIC fractionation is relatively more than that via SCX when 1 mg of protein was used as the starting material for both approaches. However, the difference in the protein product identification between ERLIC and SCX may be due to the different numbers of fractions collected from ERLIC (25 fractions) and SCX (10 fractions), and as well as the different amounts of starting materials that went on each column.

ERLIC approach in enriching for peptides for MS-based identification has been reported to provide more advantages than the SCX approach (Gan et al., 2008). As the benefit by operating in a single-step for peptide fractionation with the potential of being incorporated into high-throughput automated processes, ERLIC has been shown higher efficiency at identifying low abundant peptides and provides better coverage of peptides with acidophilic motifs (Gan et al., 2008). In this project,

my data reveals that ERLIC and SCX approaches exploit different strategies for enrichment that share some overlaps of different subsets of peptides. Therefore, complementary operation of both approaches provides more comprehensive coverage of peptides.

SDS-PAGE gel LC-MS/MS approach

The efficiency of these peptide enrichment strategies performed in my experiments was analysed by comparing the identification of protein products after LC-MS/MS analysis. From the resulting MS/MS data, 1515 protein products on average yielded from ERLIC approach while only 885 protein products on average were obtained from the SDS-PAGE gel approach (Figure 3.10) when 1 mg of protein was used as the starting material for both approaches. In the comparison between SCX and SDS-PAGE approach, 1008 protein products in average yielded from SCX approach while only 885 protein products in average yielded from SDS-PAGE gel approach (Figure 3.11) when 100 μg of protein was used as the starting material for both approaches. Overall, from the MS/MS resulting data, both ERLIC and SCX approaches resulted more protein products from the same amount of starting material than that from SDS-PAGE gel approach.

Since numerous variables were encountered during sample preparation such as gel excision procedures, and peptide purification skills, as well as that current MS analysis has not matured enough to reflect the actual effectiveness of protein enrichment strategies, these would be the major difficulties in sPEP identification.

sPEP identification and characterization

In this study, I identified eight sORFs and 11 distinct sPEPs ofthese 11 sPEPs, three have been confirmed as novel sPEPs (Table 3.2). To confirm the validity of these sPEPs, the scores in MASCOT and ProteinPilot^{TM}, and the MS/MS spectrum in b- and/or y-ion coverage. In addition,

these sPEPs should sit out of the mCDS of a gene, determined by conducting tBLASTn (NCBI) searches. These sPEPs were then validated manually in the MS/MS spectra for the final confirmation of sPEP identification.

Experiments had been performed repeatedly with minor changes in the protocol, including cell culturing and protein enrichment strategies to reconfirm those sPEPs found in previous experiments and also to obtain more data for sPEP identification. Results showed that some of those 11 sPEPs have appeared multiple times in different cell batches used throughout the experiments. This result increases the confidence of sPEP confirmation and identification.

Two recent proteomic studies claimed to find several hundreds of sPEPs (Slavoff et al., 2013, Vanderperre et al., 2013) while only 11 sPEPs were detected in my project. Reasons for the large difference in the number of sPEP identifications between mine and other researchers were evaluated by the different materials and methods used in these studies. Firstly, several aliquots of 1×10^9 of starting materials with various types of human cell lines, tissues, and fluids were used in these studies. For example, in Slavoff's group, HEK293T, HeLa, K562, COS7, and MEF cells were lysed to process polypeptides (Slavoff et al., 2013). Vanderperre *et al*. operated with more variety of starting materials by using HEK293, HeLa, human colon CCL227, CCL228, CCL233, CRL1459, and HCT116 cells, as well as human cancerous ovarian, normal ovarian, cancerous fallopian tube, normal endometrial, and lung tissues. They had also reviewed raw data published and made available by PeptideAtlas online repository (Desiere et al., 2006) from cerebrospinal fluid, plasma, and serum. They have been thorough in scrutinizing the data and observed that out-of-frame clones representing oORFs were mistakenly rejected as false positive in cDNA screening assays. In comparison, due to the limitation of the starting material, only several batches of 1×10^8 of HEK293 and HeLa cells were used in my project. Secondly, the methodology performed in my project was similar to that used in Slavoff *et al.* by using MWCO filters, followed by ERLIC

fractionation prior to MS/MS analysis. SDS-PAGE gel electrophoresis for protein separation was demonstrated in both studies and my project. SCX approach was performed as an alternative in my project. Based on the same amount of starting material, ERLIC resulted in higher efficiency at identifying low abundant peptides and provided better coverage of peptides among the approaches used in my project. In this project, results showed that ERLIC and SCX approaches exploit different strategies for enrichment that share some overlaps of different subsets of peptides. Therefore, complementary operation of both approaches provides more comprehensive coverage of peptides. All the 11 sORFs identified in my project were found through ERLIC approach and some of them were also found through SCX approach. In SDS-PAGE gel approach, none of sPEPs were identified.

Although three novel sPEPs have been identified in my work, their role in gene expression and cellular function remains unknown. Further characterisation of these sPEPs would be beneficial to explore their potential in gene regulation.

Conclusion and Future Directions

Only a few endogenous sPEPs have been detected in human cell lines from MS studies suggesting low expression of sPEPs in those cells. The dynamic range and the complexity of the cellular proteome, such as post-translational modifications create challenges when sampling low abundant and small proteins (Nielsen et al., 2006). Therefore, efficient peptide enrichment strategies are important for detecting small proteins such as sPEPs with predicted low abundance.

Moreover, the lack of information on sPEPs (eg. sub-cellular localization of the sPEPs, cell types expressed in) limits the ability to design an effective strategy that would be able to increase sensitivity of sPEP detection by LC-MS/MS. On the other hand, the limit in the detection of sPEPs may be due to MS/MS settings during analysis. The mass spectrometer was set to only find double

or triple charged peptides greater than m/z 350, which meant that only peptides of mass greater than 700 Da could be detected. Therefore, many peptides were simply never observed. In addiion, due to algorithm limits in MASCOT (ie. “false negative” results in MASCOT and also the limit to check PTM), the MS/MS data were analysed through ProteinPilot™ against RefSeq Human RNA database and cross-checked with HaltORF database.

In regard to the issue of MS analysis using MASCOT and ProteinPilot™, the y-ion coverage of peptide peaks often do not reach the maximum unless multiple analysis is attempted (Chong et al., 2006). Therefore, samples from different cell batches were used and analysed by LC-MS/MS to increase the sequence coverage. Experiments had been performed repeatedly with minor changes in the protocol, including cell culturing and protein enrichment strategies to reconfirm those sPEPs found in previous experiments and also to obtain more data for sPEP identification. Results showed that some of those 11 sPEPs have appeared few times in different cell batches throughout the experiments. This result increases the confidence of sPEP confirmation and identification.

A recent study reported that proteins of size 151-250 amino acids had higher frequency distribution than proteins of size 51-150 amino acids and even more than proteins of size $\leq$100 amino acids in eukaryotic genomes, indicating that it would be more difficult to predict biological roles in smaller proteins than larger ones although small proteins could have important biological functions (Tiessen et al., 2012).

Ribosome profiling strategies have recently emerged as a powerful tool to map which mRNA transcripts are translated at any particular stage and at what efficiency (Kuersten et al., 2013). However, proteomic studies reporting post-translational regulation of proteins, protein modifications, protein isoforms and variability within populations indicate that the current paradigm is far from accurate (Cox and Mann, 2011). Thus, to overcome this issue is to describe this diversity

by integrating data for the various parameters that can be measured, although this is computationally challenging.

Further direction to the project will be focused on chracterisation and in-cell localisation of these sPEPs to examine whether they are functional or whether they are simply unavoidable by-products of sORF *cis*-acting activities in cellular biology. The strategy designed for determining their functions is to overexpress candidate sPEPs in transfected cell lines or in whole organisms and monitor changes in phenotypes (Andrews and Rothnagel, 2014).

In summary, eight sORFs and 11 distinct sPEPs encoded by these sORFs were identified through LC-MS/MS analysis in this project, and three of these have been confirmed as novel sPEPs. This finding supports the hypothesis that a fraction of sORFs encode functional peptides that are endogenously expressed as part of the cellular proteome. The identification of the uPEP in HeLa cells also confirmed the previous finding of the same uPEP in a different human cell line in our lab. The peptide enrichment strategy performed in this project revealed that both ERLIC and SCX approaches resulted more protein products than that from SDS-PAGE gel approach. In addition, results in this project showed that protein products identification via ERLIC was ~50% higher than that via SCX. However, this may not be regarded as reliable since the starting materials that went on each column were different, as well as to the different numbers of fractions collected from ERLIC (25 fractions) and SCX (10 fractions). Overall, both ERLIC and SCX approaches gave great results in peptide separation in this project. Further characterisation of these sPEPs is necessary and beneficial to explore their potential in gene regulation.

References

ALATORRE-COBOS, F., CRUZ-RAMÍREZ, A., HAYDEN, C. A., PÉREZ-TORRES, C.-A., CHAUVIN, A.-L., IBARRA-LACLETTE, E., ALVA-CORTÉS, E., JORGENSEN, R. A. & HERRERA-ESTRELLA, L. 2012. Translational regulation of Arabidopsis XIPOTL1 is modulated by phosphocholine levels via the phylogenetically conserved upstream open reading frame 30. *Journal of experimental botany,* 63**,** 5203-5221.

ANDREWS, S. J. & ROTHNAGEL, J. A. 2014. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics,* 15**,** 193-204.

AO-KONDO, H., KOZUKA-HATA, H. & OYAMA, M. 2011. *Emergence of the Diversified Short ORFeome by Mass Spectrometry-Based Proteomics*.

BARBOSA, C., PEIXEIRO, I. & ROMÃO, L. 2013. Gene expression regulation by upstream open reading frames and human disease. *PLoS genetics,* 9**,** e1003529.

BASRAI, M. A., HIETER, P. & BOEKE, J. D. 1997. Small open reading frames: beautiful needles in the haystack. *Genome research,* 7**,** 768-771.

BERGERON, D., LAPOINTE, C., BISSONNETTE, C., TREMBLAY, G., MOTARD, J. & ROUCOU, X. 2013. An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *Journal of Biological Chemistry,* 288**,** 21824-21835.

CASSON, S. A., CHILLEY, P. M., TOPPING, J. F., EVANS, I. M., SOUTER, M. A. & LINDSEY, K. 2002. The POLARIS gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning. *The Plant Cell Online,* 14**,** 1705-1721.

CASTELLANA, N. E., PAYNE, S. H., SHEN, Z., STANKE, M., BAFNA, V. & BRIGGS, S. P. 2008. Discovery and revision of Arabidopsis genes by proteogenomics. *Proceedings of the National Academy of Sciences,* 105**,** 21034-21038.

CAZZOLA, M. & SKODA, R. C. 2000. Translational pathophysiology: a novel molecular mechanism of human disease. *Blood,* 95**,** 3280-3288.

CHEN, J. Y. & LONARDI, S. 2009. *Biological data mining*, CRC Press.

CHONG, P. K., GAN, C. S., PHAM, T. K. & WRIGHT, P. C. 2006. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections. *Journal of proteome research,* 5**,** 1232-1240.

CHUNG, W.-Y., WADHAWAN, S., SZKLARCZYK, R., POND, S. K. & NEKRUTENKO, A. 2007. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS computational biology,* 3**,** e91.

CLAVERIE, J.-M. 1997. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics,* 6**,** 1735-1744.

COLOMBANI, J., ANDERSEN, D. S. & LÉOPOLD, P. 2012. Secreted peptide Dilp8 coordinates Drosophila tissue growth with developmental timing. *Science,* 336**,** 582-585.

COX, J. & MANN, M. 2011. Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry,* 80**,** 273-299.

CROWE, M., WANG, X.-Q. & ROTHNAGEL, J. 2006. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *Bmc Genomics,* 7**,** 16.

CVIJOVIĆ, M., DALEVI, D., BILSLAND, E., KEMP, G. J. & SUNNERHAGEN, P. 2007. Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. *BMC bioinformatics,* 8**,** 295.

DAVULURI, R. V., SUZUKI, Y., SUGANO, S. & ZHANG, M. Q. 2000. CART classification of human 5′ UTR sequences. *Genome research,* 10**,** 1807-1816.

DESIERE, F., DEUTSCH, E. W., KING, N. L., NESVIZHSKII, A. I., MALLICK, P., ENG, J., CHEN, S., EDDES, J., LOEVENICH, S. N. & AEBERSOLD, R. 2006. The peptideatlas project. *Nucleic acids research,* 34**,** D655-D658.

DIBA, F., WATSON, C. S. & GAMETCHU, B. 2001. 5′ UTR sequences of the glucocorticoid receptor 1A transcript encode a peptide associated with translational regulation of the glucocorticoid receptor. *Journal of cellular biochemistry,* 81**,** 149-161.

FRANK, M. J. & SMITH, L. G. 2002. A small, novel protein highly conserved in plants and animals promotes the polarized growth and division of maize leaf epidermal cells. *Current biology,* 12**,** 849-853.

FRITH, M. C., FORREST, A. R., NOURBAKHSH, E., PANG, K. C., KAI, C., KAWAI, J., CARNINCI, P., HAYASHIZAKI, Y., BAILEY, T. L. & GRIMMOND, S. M. 2006. The abundance of short proteins in the mammalian proteome. *PLoS genetics,* 2**,** e52.

FRITSCH, C., HERRMANN, A., NOTHNAGEL, M., SZAFRANSKI, K., HUSE, K., SCHUMANN, F., SCHREIBER, S., PLATZER, M., KRAWCZAK, M. & HAMPE, J. 2012. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome research,* 22**,** 2208-2218.

GAN, C. S., GUO, T., ZHANG, H., LIM, S. K. & SZE, S. K. 2008. A comparative study of electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) versus SCX-IMAC-based methods for phosphopeptide isolation/enrichment. *Journal of proteome research,* 7**,** 4869-4877.

GELMAN, J. S., SIRONI, J., BEREZNIUK, I., DASGUPTA, S., CASTRO, L. M., GOZZO, F. C., FERRO, E. S. & FRICKER, L. D. 2013. Alterations of the intracellular peptidome in response to the proteasome inhibitor bortezomib. *PloS one,* 8**,** e53263.

HANADA, K., HIGUCHI-TAKEUCHI, M., OKAMOTO, M., YOSHIZUMI, T., SHIMIZU, M., NAKAMINAMI, K., NISHI, R., OHASHI, C., IIDA, K. & TANAKA, M. 2013. Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proceedings of the National Academy of Sciences,* 110**,** 2395-2400.

HANADA, K., ZHANG, X., BOREVITZ, J. O., LI, W.-H. & SHIU, S.-H. 2007. A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome research,* 17**,** 632-640.

HANFREY, C., ELLIOTT, K. A., FRANCESCHETTI, M., MAYER, M. J., ILLINGWORTH, C. & MICHAEL, A. J. 2005. A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. *Journal of Biological Chemistry,* 280**,** 39229-39237.

HANYU-NAKAMURA, K., SONOBE-NOJIMA, H., TANIGAWA, A., LASKO, P. & NAKAMURA, A. 2008. Drosophila Pgc protein inhibits P-TEFb recruitment to chromatin in primordial germ cells. *Nature,* 451**,** 730-733.

HAYDEN, C. A. & BOSCO, G. 2008. Comparative genomic analysis of novel conserved peptide upstream open reading frames in Drosophila melanogaster and other dipteran species. *Bmc Genomics,* 9**,** 61.

HAYDEN, C. A. & JORGENSEN, R. A. 2007. Identification of novel conserved peptide uORF homology groups in Arabidopsis and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. *BMC biology,* 5**,** 32.

IACONO, M., MIGNONE, F. & PESOLE, G. 2005. uAUG and uORFs in human and rodent 5′ untranslated mRNAs. *Gene,* 349**,** 97-105.

INGOLIA, N. T., LAREAU, L. F. & WEISSMAN, J. S. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell,* 147**,** 789-802.

JORGENSEN, R. A. & DORANTES-ACOSTA, A. E. 2012. Conserved peptide upstream open reading frames are associated with regulatory genes in angiosperms. *Frontiers in plant science,* 3.

KAGEYAMA, Y., KONDO, T. & HASHIMOTO, Y. 2011. Coding vs non-coding: Translatability of short ORFs found in putative non-coding transcripts. *Biochimie,* 93**,** 1981-1986.

KASTENMAYER, J. P., NI, L., CHU, A., KITCHEN, L. E., AU, W.-C., YANG, H., CARTER, C. D., WHEELER, D., DAVIS, R. W. & BOEKE, J. D. 2006. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome research,* 16**,** 365-373.

KONDO, T., PLAZA, S., ZANET, J., BENRABAH, E., VALENTI, P., HASHIMOTO, Y., KOBAYASHI, S., PAYRE, F. & KAGEYAMA, Y. 2010. Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. *Science,* 329**,** 336-339.

KOZAK, M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic acids research,* 15**,** 8125-8148.

KUERSTEN, S., RADEK, A., VOGEL, C. & PENALVA, L. O. 2013. Translation regulation gets its ‘omics’ moment. *Wiley Interdisciplinary Reviews: RNA,* 4**,** 617-630.

LADOUKAKIS, E., PEREIRA, V., MAGNY, E. G., EYRE-WALKER, A. & COUSO, J. P. 2011. Hundreds of putatively functional small open reading frames in Drosophila. *Genome biology,* 12**,** R118.

LAW, G. L., RANEY, A., HEUSNER, C. & MORRIS, D. R. 2001. Polyamine regulation of ribosome pausing at the upstream open reading frame of S-adenosylmethionine decarboxylase. *Journal of Biological Chemistry,* 276**,** 38036-38043.

LEASE, K. A. & WALKER, J. C. 2006. The Arabidopsis unannotated secreted peptide database, a resource for plant peptidomics. *Plant physiology,* 142**,** 831-838.

LEE, S., LIU, B., LEE, S., HUANG, S.-X., SHEN, B. & QIAN, S.-B. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences,* 109**,** E2424-E2432.

MAGNY, E. G., PUEYO, J. I., PEARL, F. M., CESPEDES, M. A., NIVEN, J. E., BISHOP, S. A. & COUSO, J. P. 2013. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science,* 341**,** 1116-1120.

MARINO, S. M. & GLADYSHEV, V. N. 2012. Analysis and functional prediction of reactive cysteine residues. *Journal of Biological Chemistry,* 287**,** 4419-4425.

MCMURRAY, C. T. 2010. Mechanisms of trinucleotide repeat instability during human development. *Nature Reviews Genetics,* 11**,** 786-799.

MEIJER, H. & THOMAS, A. 2002. Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. *Biochem. J,* 367**,** 1-11.

MERCER, T. R., WILHELM, D., DINGER, M. E., SOLDÀ, G., KORBIE, D. J., GLAZOV, E. A., TRUONG, V., SCHWENKE, M., SIMONS, C. & MATTHAEI, K. I. 2011. Expression of distinct RNAs from 3′ untranslated regions. *Nucleic acids research,* 39**,** 2393-2403.

MICHEL, A. M., CHOUDHURY, K. R., FIRTH, A. E., INGOLIA, N. T., ATKINS, J. F. & BARANOV, P. V. 2012. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome research,* 22**,** 2219-2229.

NARITA, N. N., MOORE, S., HORIGUCHI, G., KUBO, M., DEMURA, T., FUKUDA, H., GOODRICH, J. & TSUKAYA, H. 2004. Overexpression of a novel small peptide ROTUNDIFOLIA4 decreases cell proliferation and alters leaf shape in Arabidopsis thaliana. *The Plant Journal,* 38**,** 699-713.

NEAFSEY, D. E. & GALAGAN, J. E. 2007. Dual modes of natural selection on upstream open reading frames. *Molecular biology and evolution,* 24**,** 1744-1751.

NGUYEN, H. L., YANG, X. & OMIECINSKI, C. J. 2013. Expression of a novel mRNA transcript for human microsomal epoxide hydrolase (EPHX1) is regulated by short open reading frames within its 5′-untranslated region. *rna,* 19**,** 752-766.

NIELSEN, M. L., SAVITSKI, M. M. & ZUBAREV, R. A. 2006. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Molecular & Cellular Proteomics,* 5**,** 2384-2391.

NORMARK, S., BERGSTROM, S., EDLUND, T., GRUNDSTROM, T., JAURIN, B., LINDBERG, F. P. & OLSSON, O. 1983. Overlapping genes. *Annual review of genetics,* 17**,** 499-525.

ORR, H. T. & ZOGHBI, H. Y. 2007. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.,* 30**,** 575-621.

OYAMA, M., ITAGAKI, C., HATA, H., SUZUKI, Y., IZUMI, T., NATSUME, T., ISOBE, T. & SUGANO, S. 2004. Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome research,* 14**,** 2048-2052.

OYAMA, M., KOZUKA-HATA, H., SUZUKI, Y., SEMBA, K., YAMAMOTO, T. & SUGANO, S. 2007. Diversity of translation start sites may define increased complexity of the human short ORFeome. *Molecular & Cellular Proteomics,* 6**,** 1000-1006.

PENAGARIKANO, O., MULLE, J. G. & WARREN, S. T. 2007. The pathophysiology of fragile x syndrome. *Annu. Rev. Genomics Hum. Genet.,* 8**,** 109-129.

PENDLETON, L. C., GOODWIN, B. L., SOLOMONSON, L. P. & EICHLER, D. C. 2005. Regulation of endothelial argininosuccinate synthase expression and NO production by an upstream open reading frame. *Journal of Biological Chemistry,* 280**,** 24252-24260.

PESOLE, G., GISSI, C., GRILLO, G., LICCIULLI, F., LIUNI, S. & SACCONE, C. 2000. Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. *Gene,* 261**,** 85-91.

QURASHI, A., SAHIN, H. B., CARRERA, P., GAUTREAU, A., SCHENCK, A. & GIANGRANDE, A. 2007. HSPC300 and its role in neuronal connectivity. *Neural development,* 2**,** 18.

RIBRIOUX, S., BRÜNGGER, A., BAUMGARTEN, B., SEUWEN, K. & JOHN, M. R. 2008. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *Bmc Genomics,* 9**,** 122.

ROGOZIN, I. B., KOCHETOV, A. V., KONDRASHOV, F. A., KOONIN, E. V. & MILANESI, L. 2001. Presence of ATG triplets in 5′ untranslated regions of eukaryotic cDNAs correlates with a 'weak'context of the start codon. *Bioinformatics,* 17**,** 890-900.

RÖHRIG, H., SCHMIDT, J., MIKLASHEVICHS, E., SCHELL, J. & JOHN, M. 2002. Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proceedings of the National Academy of Sciences,* 99**,** 1915-1920.

RONSIN, C., CHUNG-SCOTT, V., POULLION, I., AKNOUCHE, N., GAUDIN, C. & TRIEBEL, F. 1999. A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ. *The Journal of Immunology,* 163**,** 483-490.

SAVARD, J., MARQUES-SOUZA, H., ARANDA, M. & TAUTZ, D. 2006. A Segmentation Gene in< i> Tribolium</i> Produces a Polycistronic mRNA that Codes for Multiple Conserved Peptides. *Cell,* 126**,** 559-569.

SCHÄGGER, H. 2006. Tricine–SDS-PAGE. *Nature protocols,* 1**,** 16-22.

SETHURAMAN, M., MCCOMB, M. E., HUANG, H., HUANG, S., HEIBECK, T., COSTELLO, C. E. & COHEN, R. A. 2004. Isotope-coded affinity tag (ICAT) approach to redox proteomics: identification and quantitation of oxidant-sensitive cysteine thiols in complex protein mixtures. *Journal of proteome research,* 3**,** 1228-1233.

SKARSHEWSKI, A., STANTON-COOK, M., HUBER, T., AL MANSOORI, S., SMITH, R., BEATSON, S. A. & ROTHNAGEL, J. A. 2014. uPEPperoni: An online tool for upstream open reading frame location and analysis of transcript conservation. *BMC bioinformatics,* 15**,** 36.

SLAVOFF, S. A., MITCHELL, A. J., SCHWAID, A. G., CABILI, M. N., MA, J., LEVIN, J. Z., KARGER, A. D., BUDNIK, B. A., RINN, J. L. & SAGHATELIAN, A. 2013. Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature chemical biology,* 9**,** 59-64.

SMEEKENS, S., MA, J., HANSON, J. & ROLLAND, F. 2010. Sugar signals and molecular networks controlling plant growth. *Current opinion in plant biology,* 13**,** 273-278.

SUBIRANA, J. A. & MESSEGUER, X. 2008. Structural families of genomic microsatellites. *Gene,* 408**,** 124-132.

SZE, S., HAO, P., ZHANG, H., GUO, T., LI, X., YANG, J., TAM, J., LIM, S. & ALPERT, A. 2010. ERLIC and Proteomics: Simultaneous Isolation of Phospho-and Glycopeptides and Superior Fractionation of Complex Tryptic Digests. *Journal of biomolecular techniques: JBT,* 21**,** S32.

TAKAHASHI, H., TAKAHASHI, A., NAITO, S. & ONOUCHI, H. 2012. BAIUCAS: a novel BLAST-based algorithm for the identification of upstream open reading frames with conserved amino acid sequences and its application to the Arabidopsis thaliana genome. *Bioinformatics,* 28**,** 2231-2241.

TIESSEN, A., PÉREZ-RODRÍGUEZ, P. & DELAYE-ARREDONDO, L. J. 2012. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC research notes,* 5**,** 85.

TINOCO, A. D., TAGORE, D. M. & SAGHATELIAN, A. 2010. Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform. *Journal of the American Chemical Society,* 132**,** 3819-3830.

TRAN, M. K., SCHULTZ, C. J. & BAUMANN, U. 2008. Conserved upstream open reading frames in higher plants. *Bmc Genomics,* 9**,** 361.

VANDERPERRE, B., LUCIER, J.-F., BISSONNETTE, C., MOTARD, J., TREMBLAY, G., VANDERPERRE, S., WISZTORSKI, M., SALZET, M., BOISVERT, F.-M. & ROUCOU, X. 2013. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS one,* 8**,** e70698.

VANDERPERRE, B., LUCIER, J.-F. & ROUCOU, X. 2012a. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. *Database: The Journal of Biological Databases and Curation,* 2012.

VANDERPERRE, B., LUCIER, J.-F. & ROUCOU, X. 2012b. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. *Database,* 2012**,** bas025.

VANDERPERRE, B., STASKEVICIUS, A. B., TREMBLAY, G., MCCOY, M., O'NEILL, M. A., CASHMAN, N. R. & ROUCOU, X. 2011. An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *The FASEB Journal,* 25**,** 2373-2386.

VAUGHN, J. N., ELLINGSON, S. R., MIGNONE, F. & VON ARNIM, A. 2012. Known and novel post-transcriptional regulatory sequences are conserved across plant families. *RNA,* 18**,** 368-384.

WANG, R.-F., PARKHURST, M. R., KAWAKAMI, Y., ROBBINS, P. F. & ROSENBERG, S. A. 1996. Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *The Journal of experimental medicine,* 183**,** 1131-1140.

WANG, X. Q. & ROTHNAGEL, J. A. 2004. 5′-Untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic acids research,* 32**,** 1382-1391.

WEIST, S., ERAVCI, M., BROEDEL, O., FUXIUS, S., ERAVCI, S. & BAUMGARTNER, A. 2008. Results and reliability of protein quantification for two-dimensional gel electrophoresis strongly depend on the type of protein sample and the method employed. *Proteomics,* 8**,** 3389-3396.

WEN, Y., LIU, Y., XU, Y., ZHAO, Y., HUA, R., WANG, K., SUN, M., LI, Y., YANG, S. & ZHANG, X.-J. 2009. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nature genetics,* 41**,** 228-233.

WERNER, M., FELLER, A., MESSENGUY, F. & PIÉRARD, A. 1987. The leader peptide of yeast gene< i> CPA1</i> is essential for the translational repression of its expression. *Cell,* 49**,** 805-813.

WETHMAR, K., SMINK, J. J. & LEUTZ, A. 2010. Upstream open reading frames: molecular switches in (patho) physiology. *Bioessays,* 32**,** 885-893.

WILLIS, A. E. 1999. Translational control of growth factor and proto-oncogene expression. *The international journal of biochemistry & cell biology,* 31**,** 73-86.

WU, H., MA, B.-G., ZHAO, J.-T. & ZHANG, H.-Y. 2007. How similar are amino acid mutations in human genetic diseases and evolution. *Biochemical and biophysical research communications,* 362**,** 233-237.

XU, H., WANG, P., FU, Y., ZHENG, Y., TANG, Q., SI, L., YOU, J., ZHANG, Z., ZHU, Y. & ZHOU, L. 2010. Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell research,* 20**,** 445-457.

YANG, X., TSCHAPLINSKI, T. J., HURST, G. B., JAWDY, S., ABRAHAM, P. E., LANKFORD, P. K., ADAMS, R. M., SHAH, M. B., HETTICH, R. L. & LINDQUIST, E. 2011. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome research,* 21**,** 634-641.

ZAREI, M., SPRENGER, A., GRETZMEIER, C. & DENGJEL, J. 2012. Combinatorial use of electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) and strong cation exchange (SCX) chromatography for in-depth phosphoproteome analysis. *Journal of proteome research,* 11**,** 4269-4276.

ZAREI, M., SPRENGER, A., METZGER, F., GRETZMEIER, C. & DENGJEL, J. 2011. Comparison of ERLIC–TiO2, HILIC–TiO2, and SCX–TiO2 for global phosphoproteomics approaches. *Journal of proteome research,* 10**,** 3474-3483.

ZHANG, Z. & DIETRICH, F. S. 2005. Identification and characterization of upstream open reading frames (uORF) in the 5′ untranslated regions (UTR) of genes in Saccharomyces cerevisiae. *Current genetics,* 48**,** 77-87.

Appendices

Appendix 1

Table A1. Conservation of uORFs in difference species.

Species with uORF sequence conservation	Number of conserved uORFs
Human and mice (Crowe et al., 2006)	247
Arabidopsis Thaliana and rice (Hayden and Jorgensen, 2007)	15
Rice, sorghum, wheat, maize, and barley (Tran et al., 2008)	29
Arabidopsis, grapevine, tobacco, soybeans, orange, and cotton (Vaughn et al., 2012)	18
Cryptococcus neoformans (Neafsey and Galagan, 2007)	122
Drosophila melanogaster and other dipteran EST sequences (Hayden and Bosco, 2008)	44
Human (Slavoff et al., 2013)	15

Table A2. Conservation of sORFs in bioinformatics studies.

Reference	Species involved	Number of putative coding sORFs	Number of transcripts analyzed
uORFs			
(Iacono et al., 2005)	Conservation of human and mouse uORFs.	43	27,660 (human)
(Crowe et al., 2006)	Conserved uORFs between human and mouse; bias towards optimal Kozak sequences; evidence of purifying selection.	204	21,768 (human)
(Hayden and Bosco, 2008)	Conserved uORFs between *D. melanogaster, A. gambiae* and EST sequences from *D. simulans*, *D. yakuba*, *D. erecta*, and *D. ananassae*; evidence of purifying selection.	44	19,389 (*D. melanogaster*)
(Hayden and Jorgensen, 2007)	Conserved uORFs between *Arabidopsis* and rice; four uORFs had conservation between others species; evidence of purifying selection.	19	34,000 (*Arabidopsis*)
(Tran et al., 2008)	Conservation of uORFs between Rice, wheat, barley, maize, sorghum, and	29	32,127 (rice)

	Arabidopsis.	15	
(Takahashi et al., 2012)	Conserved uORFs in *Arabidopsis* based on analysis of EST sequences; evidence of purifying selection.	18	27,101 (*Arabidopsis*)
(Vaughn et al., 2012)	Conserved uORFs between *Arabidopsis*, tobacco, grapevine, soybeans, orange and cotton; evidence of purifying selection.	18	10,122 (*Arabidopsis*)
(Zhang and Dietrich, 2005)	Conserved uORFs between *S. cerevisiae and S. paradoxus, S. bayanus, S. mikatae, S. kudriavzevii, S. kluyveri, S. castelii, A. gossypii, C. glabrata , K. lactis and K. waltii;* evidence of transcription.	19	5,542 (*S. cerevisiae)*
(Cvijović et al., 2007)	Conserved uORFs between *S. cerevisiae* and at least one of the following fungual species; *S. paradoxus, S. mikatae, S. bayanus, S. kudriavzevii, S, castellii or S. kluyveri.*	379	5,602 (*S, cerevisiae)*
(Neafsey and Galagan, 2007)	Conserved uORFs between the four *Cryptococcus neoformans* strains (JEC21, TIGR, WM276, and H99); evidence of purifying selection in 12 uORFs.	122	2,167 (*C. neoformans*)

ncRNAs			
(Frith et al., 2006)	sORFs in *Mus musculus* using CRITCA; evidence of conservation in rats and humans; evidence of transcription; evidence of localization.	1,240	102,801 transcripts 40,841 sORFs (mouse)
(Ladoukakis et al., 2011)	Conserved ncRNAs between D. melanogaster and *D. pseudoobscura*; evidence of transcription; evidence of purifying selection; evidence of syntenic conservation.	401	593,586 sORFs (*D. melanogaster)*
(Hanada et al., 2007)	ncRNAs in *Arabidopsis;* evidence of transcription; evidence of conservation in *B. oleracea, O. sativa, P. trichocarpa, M. truncatula* or *L. corniculatus;* evidence of purifying selection.	3,241	570,948 sORFs (*Arabidopsis*)
(Hanada et al., 2013)	ncRNAs in *Arabidopsis;* evidence of transcription; evidence of conservation in *P. patens, M. moellendorffii, Z. mays, S. bicolor, B. distachyon, O. sativa, M. guttatus, V. vinifera, R. communis, M. esculenta, P. trichocarpa, C. sativus, G. max, M. truncatula, C. papaya, or A.*	2,302	96,358 transcripts (*Arabidopsis*)

	lyrata; evidence of purifying selection.		
(Lease and Walker, 2006)	ncRNAs conserved between Arabidopsis and rice; evidence of transcription; evidence of clustering of potential gene families.	1,044	606,285 sORFs (*Arabidopsis*)
(Yang et al., 2011)	ncRNAs in *P. deltoids*; evidence of conservation; evidence of coding potential; evidence of clustered peptide families; evidence of protein domains/motifs and evidence of expression via mass spectrometry.	1,469	~2.6 million ESTs 12,852 sORFs (*P. deltoids)*
(Kastenmayer et al., 2006)	Characterised ncRNAs in *Saccharomyces cerevisiae* previously identified as putative coding sORFs based on evidence of transcription or translation. Further of evidence of conservation in *S. pombe*, *C. elegans*, *A. thaliana*, *D. melanogaster*, *M. musculus* or *H. sapiens*; evidence of functionality based on gene deletion experiments.	299	N/A
oORFS			

(Chung et al., 2007)	Conserved oORFs between human, mice and rat or dog that are under purifying selection.	40	14159 (human)
(Ribrioux et al., 2008)	Conserved oORFs between human, mice and rat that contained Kozak sequences.	215	9163 (human)
(Xu et al., 2010)	Conserved oORFs between human and mice.	168	26009 (human)
(Vanderperre et al., 2012a)	oORFs on alternate reading frames from human mRNA that had strong Kozak sequences.	24,547	~76000 (human)

Table A3. Functional sPEPs.

Species	Genes or transcripts	Number of residues in sPEP	Function
Upstream sPEPs			
Arabidopsis thaliana	*GBF6* (Smeekens et al., 2010)	28	Expression of the CDS is modulated by sucrose levels through a conserved sPEP
	SAMDC (Hanfrey et al., 2005)	52	Expression of the CDS is regulated by polyamines binding to the nascent upstream sPEP; orthologous to human *SAMDC*
	XPL1 (Alatorre-Cobos et al., 2012)	26	Expression of the CDS is regulated by phosphocholine binding to the sPEP
Saccharomyces cerevisiae	*CPA1* (Werner et al., 1987)	25	The sPEP reduces expression of the CDS through ribosomal stalling and blocking translation in response to increased arginine levels
Humans	*ASS1* (Pendleton et al., 2005)	44	The sPEP regulates expression of *ASS1* in a *trans*-suppressive manner
	EPHX1 (Nguyen et al., 2013)	17 & 26	Expression of *EPHX1* is inhibited by *trans*-acting sPEPs that are encoded by two uORFs through interactions with the translation machinery
	HR (Wen et al., 2009)	34	The sPEP is implicated in the regulation of *HR*; 13 causative mutations of Marie Unna hereditary hypotrichosis have been identified within the second uORF
	MKKS (Akimoto et al, 2013)	63 & 50	Both sPEPs localize to the mitochondrial membrane and are predicted to function independently of MKKS
	NR3C1 (Diba et al., 2001)	93	The sPEP localizes to the cell membrane and regulates expression of the glucocorticoid receptor in a *trans*-acting manner through interaction with unknown cellular factors
	SAMDC (Law et al., 2001)	6	Expression of the CDS is regulated by polyamines binding to the nascent upstream sPEP; orthologous to *A. thaliana SAMDC*
Intergenic sPEPs			

A. thaliana	*PLS* (Casson et al., 2002)	36	The sPEP is required for correct auxin–cytokinin homeostasis to modulate root growth and leaf vascular patterning
	ROT4 (Narita et al., 2004)	53	The sPEP is involved in regulation of leaf shape by reducing cell proliferation in lateral organs
Drosophila melanogaster	*Ilp8* (Colombani et al., 2012)	150	The sPEP provides a signal that promotes the delay of metamorphosis in response to conditions that that alter growth in imaginal discs
	HSPC300 (Qurashi et al., 2007)	75	The sPEP is a component of the WAVE–SCAR complex and is important in nervous system development for axonogenesis and neuromuscular synapse morphogenesis; *HSPC300* is orthologous to *brk1*
	pgc (Hanyu-Nakamura et al., 2008)	71	The sPEP is essential for repressing Ser2 phosphorylation in the carboxy-terminal domain of RNA polymerase II in newly formed pole cells (which are the early germline progenitors) and thus has a fundamental role in germ-cell specification
	tal (Kondo et al., 2010)	11 & 32	The sORFs encode three peptides of 11 residues and one peptide of 32 residues that are essential for embryonic development and that are required for formation of epithelial architecture; *tal* is orthologous to *Mlpt*
	RanGAP (Magny et al., 2013)	28 & 29	Both sPEPs are involved in the regulation of Ca2+ trafficking; alterations result in irregular muscle contractions
Maize	*brk1* (Frank and Smith, 2002)	84	The sPEP promotes multiple actin-dependent cell polarization events in the developing leaf epidermis; *brk1* is orthologous to *HSPC300*
Soybean	*ENOD40-1* (Röhrig et al., 2002)	12 & 24	The sPEP binds to nodulin 100 (which is a subunit of sucrose synthase) and is likely to be involved in the control of sucrose use in nitrogen-fixing nodules
Tribolium castaneum	*Mlpt* (Savard et al., 2006)	10, 11, 15 & 23	The sORFs encode four sPEPs with roles in embryonic development, particularly the development of abdominal segments; *Mlpt* is orthologous to *tal*
Overlapping sPEPs and downstream sPEPs			
Humans	*TYRP1* (Wang et al., 1996)	24	The sPEP is co-expressed from the *TYRP1* transcript

	CASP1 (Ronsin et al., 1999)	151	The sPEP is expressed from the intestinal carboxyl esterase gene and is recognized by human leukocyte antigen-B7-restricted renal cell carcinoma-reactive T cell clone
	AltPrP (Vanderperre et al., 2011)	73	The sPEP is co-expressed from the prion protein transcript in brain homogenates, primary neurons and peripheral blood mononuclear cells; it localizes to the mitochondria
	AltATXN1 (Bergeron et al., 2013)	185	The sPEP is co-expressed from the *ATXN1* transcript and is expressed in the cerebellum; it colocalizes and interacts with the ATXN1 protein in the nucleus
	AltMRVI1 (Vanderperre et al., 2013)	134	The sPEP colocalizes to the nucleus and interacts with BRCA1

Table A4. Parameters set for MASCOT MS/MS ion search.

Fixed modification	Variable modifications	Missed cleavages	Peptide tolerance (±)
• Carbamidomethyl (C)	• Gly->pyro-Glu (N-term Q) • Gln->pyro-Glu (N-term Q) • Oxidation (M)	• Allow up to 2	• 50 ppm

Peptide charge	Taxonomy	Enzyme digestion
• 2+, 3+, & 4+	• Homo sapiens	• Semi-trypsin

Appendix 2

Details of protocols in 2D Quant kit

Peptide samples were quantified using 2D Quant kit according to the manufacturer's specifications. 2 mg/ml Bovine serum albumin (BSA) standard solution provided with the kit was used to prepare a standard curve (0 µl, 2.5 µl, 5 µl, 7.5 µl, 10 µl and 12.5 µl). 2 µl and 15 µl of each protein sample were analysed each time. Precipitant (250 µl) was added to each tube and vortexed briefly. Tubes were then incubated for 2 to 3 minutes at room temperature. Co-precipitant (250 µl) was added to each tube followed by vortexing and centrifugation at 10, 000 x g for 5 minutes. Supernatant was then decanted and removed with pipette carefully. 50 µl copper solution and 200 µl of water were added into each tube and vortexed briefly. 500 µl working colour reagent was added into each tube. Tubes were then incubated for 20 minutes at room temperature before reading the absorbance, standard at 480nm using water as reference.

Appendix 3

Characterisation of sPEPs found in HEK cells

Conservation of sPEPs with other species

a) sPEP translated from *LINC00950* non-coding RNA

Blast result using nucleotide sequence- Blastn result:

Description	Total score	Query cover	E value	Identity
Human DNA sequence from clone RP11-112J3 on chromosome 9p13.1-13.3, complete sequence	942	100%	0.0	100%
Homo sapiens clone 23583 mRNA sequence	941	99%	0.0	100%
Homo sapiens mRNA; cDNA DKFZp547N0218 (from clone DKFZp547N0218)	935	100%	0.0	99%
PREDICTED: Cavia porcellus olfactory receptor 13C3-like (LOC100729728), mRNA	113	20%	1e-21	86%

Blast search using protein sequence- Blastp result:

Conserved Species	Identities/ E-value
Homo sapiens cartilage-hair hypoplasia region gene sequence	100% / 1e-08 Query 1 LRISNGSDEISLPLTYWPWPKCL* 72

	LRISNGSDEISLPLTYWPWPKCL* Sbjct 72937 LRISNGSDEISLPLTYWPWPKCL* 73008
PREDICTED: nesprin-1-like [Neolamprologus brichardi]	73 % / 106 Query 1 ISNGSDEISLP 11 IS GSDEI P Sbjct 2877 ISAGSDEIAFP 2887
PREDICTED: aldehyde dehydrogenase family 9 member A1-B-like [Maylandia zebra]	73 % / 191 Query 1 ISNGSDEISLP 11 IS GS EI LP Sbjct 25 ISSGSVEITLP 35
coiled-coil domain-containing protein [Cricetulus griseus]	73% / 263 Query 1 ISNGSDEISLP 11 IS SD+ISLP Sbjct 1322 IS--SDDISLP 1330
Fat4 protein, partial [Mus musculus]	80 % / 78 Query 2 SNGSDEISLP 11 S G DEISLP

	Sbjct 1531 SQGPDEISLP 1540
protocadherin Fat 4 [Rattus norvegicus]	80 %/ 78 Query 2 SNGSDEISLP 11 S G DEISLP Sbjct 4491 SQGPDEISLP 4500
Fat4 [Mus musculus]	80 % / 78 Query 2 SNGSDEISLP 11 S G DEISLP Sbjct 4493 SQGPDEISLP 4502
hypothetical protein PANDA_000833 [Ailuropoda melanoleuca]	69 % / 105 Query 1 ISNGSDEI--SLP 11 ISNGS E+ SLP Sbjct 809 ISNGSAEVDLSLP 821

b) sPEP translated from *TM9SF3*

Blast result using nucleotide sequence- Blastn result:

Description	Total score	Query cover	E value	Identity
PREDICTED: Pongo abelii transmembrane 9 superfamily member 3 (TM9SF3), mRNA	449	100%	6e-123	93%
PREDICTED: Macaca fascicularis uncharacterized LOC101926339 (LOC101926339), mRNA	363	100%	8e-97	88%
PREDICTED: Macaca mulatta transmembrane 9 superfamily member 3, transcript variant 3 (TM9SF3), mRNA	363	100%	8e-97	88%
PREDICTED: Orcinus orca transmembrane 9 superfamily member 3 (TM9SF3), mRNA	337	98%	5e-89	87%
PREDICTED: Odobenus rosmarus divergens transmembrane 9 superfamily member 3 (TM9SF3), mRNA	311	89%	3e-81	87%
Bos taurus transmembrane 9 superfamily member 3 (TM9SF3), mRNA	272	91%	1e-69	85%
PREDICTED: Canis lupus familiaris transmembrane 9 superfamily member 3 (TM9SF3), transcript variant X1, mRNA	254	82%	5e-64	85%

Blast search using protein sequence- Blastp result:

Conserved Species	**Identities/ E-value**
PREDICTED: fibroblast growth factor 3 [Bos mutus]	81 % / 0.22 Query 1 ATAAAEEAAAGPGPVR 16 A AAAEEA AGPGP R Sbjct 246 AAAAAEEA-AGPGPGR 260
PREDICTED: epiplakin, partial [Condylura cristata]	80 % / 0.32 Query 2 TAAAEEAAAGPGPVR 16 TA EEAAAGPG VR Sbjct 1021 TAIVEEAAAGPGRVR 1035
PREDICTED: cryptochrome-2 [Erinaceus europaeus]	86 % / 0.42 Query 2 TAAAEEAAAGPGPV 15 TAAA AAAGPGPV Sbjct 7 TAAATAAAAGPGPV 20
PREDICTED: LOW QUALITY PROTEIN: ubiquitin specific peptidase 35 [Bos mutus]	86 % / 0.59 Query 3 AAA--EEAAAGPGP 14 AAA EEAAAGPGP

	Sbjct 744 AAADGEEAAAGPGP 757
PREDICTED: ubiquitin carboxyl-terminal hydrolase 35 [Bubalus bubalis]	86 % 0.59 Query 3 AAA--EEAAAGPGP 14 AAA EEAAAGPGP Sbjct 728 AAADGEEAAAGPGP 741
PREDICTED: ubiquitin carboxyl-terminal hydrolase 35 [Bos taurus]	86 % / 0.58 Query 3 AAA--EEAAAGPGP 14 AAA EEAAAGPGP Sbjct 735 AAADGEEAAAGPGP 748
zinc finger protein ZIC 5 [Mus musculus]	79 % / 2.0 Query 1 ATAAAEEAAAGPGP 14 AAAA AAAGPGP Sbjct 310 AAAAAAAAAAGPGP 323
PREDICTED: MAP kinase-interacting serine/threonine-protein kinase 2 [Cavia porcellus]	91 % / 3.6 Query 4 AAEEAAAGPGP 14 AAEEAAAG GP Sbjct 413 AAEEAAAGQGP 423

PREDICTED: ubiquitin-associated domain-containing protein 1 [Macaca fascicularis]	85 % / 5.0 Query 1 ATAAAEEAAAGPG 13 ATAAA EAAAG G Sbjct 308 ATAAAPEAAAGAG 320
zinc family member 5 protein [Homo sapiens]	83 % / 9.3 Query 3 AAAEEAAAGPGP 14 AAA AAAGPGP Sbjct 314 AAAAAAAAGPGP 325
paralemin [Homo sapiens]	69 % / 12 Query 1 ATAAAEEAAAGPGPVR 16 AAAAE AAP PVR Sbjct 104 APAAAKENAAAPSPVR 119

c) sPEP translated from *PTPN21*

Blast result using nucleotide sequence- Blastn result:

Description	**Total score**	**Query cover**	**E value**	**Identity**
PREDICTED: Homo sapiens spermatogenesis associated 7 (SPATA7), transcript variant X7, mRNA	244	100%	5e-62	100%
PREDICTED: Nomascus leucogenys protein tyrosine phosphatase, non-receptor type 21 (PTPN21), mRNA	243	100%	2e-61	100%
PREDICTED: Gorilla gorilla gorilla tyrosine-protein phosphatase non-receptor type 21-like (LOC101141605), mRNA	243	100%	2e-61	100%
PREDICTED: Pan troglodytes protein tyrosine phosphatase, non-receptor type 21, transcript variant 1 (PTPN21), mRNA	243	100%	2e-61	100%

Blast search using protein sequence- Blastp result:

Conserved Species	**Identities/ E-value**
PREDICTED: protein FAM13A-like [Condylura cristata]	78 % / 83 Query 1 GTITEYLSR 9 GTI EYL R Sbjct 93 GTIVEYLTR 101
similar to F-box protein FBL2, isoform CRA_b [Rattus	75 % / 273

norvegicus]	Query 2 TITEYLS R 9 TITEY+ R Sbjct 109 TITEYMGR 116
apolipoprotein B editing complex 1, isoform CRA_b [Mus musculus]	86 % / 388 Query 3 ITEYLSR 9 ITE+LSR Sbjct 111 ITEFLSR 117

Appendix 4

a) Characterisation of sPEP translated from *LINC00950* non-coding RNA found in HEK cells

Information was obtained using a web-based peptide/protein analysis tool from LifeTein (http://lifetein.com/peptide-analysis-tool.html), indicating properties of the sPEP including protein hydrophobicity, secondary structure prediction, transmembrane region predictions, protein flexibility.

Primary information of sPEP translated from *LINC00950* non-coding RNA	
Interpreted amino acid sequence	YYYHPVPIPVPQQIFWLLNVLNDFLRISNGSDEISLPLTYWPWPKCL
Number of Standard Residues	47
Chemical Formula	$C_{274}H_{395}N_{61}O_{67}S$
Molecular Weight	5647.86 g/mol
Sequence Composition (in percentage)	Acidic: 6.38 Basic: 6.38 Neutral: 36.17 Hydrophobic: 51.06
Amino Acid Composition	C = 1 (2.13) D = 2 (4.26) E = 1 (2.13) F = 2 (4.26) G = 1 (2.13) H = 1 (2.13) I = 4 (8.51) K = 1 (2.13) L = 7 (14.89) N = 3 (6.38) P = 7 (14.89) Q = 2 (4.26) R = 1 (2.13) S = 3 (6.38) T = 1 (2.13) V = 3 (6.38) W = 3 (6.38) Y = 4 (8.51)
Isoelectric Point (pI)	5.29

Net Charge at pH 7	-0.98
Charge	0
Charge Attribute	Neutral
Secondary Structure Predicition	
Amphiphilicity Helix	Y Y Y H P V P I P V P Q Q I F W L L N V **L N D F L** R I S N G S D E I S L P L T Y W P W P K C L Legend: **Helix**
Chou-Fasman Prediction	Y Y Y **H** P V P **I P V P Q Q I F W L L N V L N D F L R I S N G S D E I S L** P L T Y **W** P**W P** K C L Legend:**Alpha**,**Beta**, **Turn**
GOR-I Prediction	Y Y Y H P V P I **P V P Q Q I F W L L N V L N D F L R I S N G S D E I S L P L T** Y W P W P K C L Legend:**Alpha**,**Beta**, **Turn**, **Coil**

b) Characterisation of sPEP translated from *TM9SF3* found in HEK cells

Information wasobtained using a web-based peptide/protein analysis tool from LifeTein (http://lifetein.com/peptide-analysis-tool.html), indicating properties of the sPEP including protein hydrophobicity, secondary structure prediction, transmembrane region predictions, protein flexibility.

Primary information of sPEP translated from *TM9SF3*	
Interpreted amino acid sequence	EEEAEEARGAGEAQERAVTATAAAEEAAAGPGPVRL SAPRGCESAAAEAAAAEEAAAGGGVGEPAPGRRGAE DEAAAWRSWRGGGRRAVAAAAAAAPDPGGRARTH VSR
Number of Standard Residues	109
Chemical Formula	$C_{435}H_{699}N_{153}O_{154}S$
Molecular Weight	10568.73 g/mol
Sequence Composition (in percentage)	Acidic: 16.51 Basic: 12.84 Neutral: 28.44 Hydrophobic: 42.2
Amino Acid Composition	A = 37 (33.94) C = 1 (0.92) D = 2 (1.83) E = 16 (14.68) G = 16 (14.68) H = 1 (0.92) L = 1 (0.92) P = 7 (6.42) Q = 1 (0.92) R = 13 (11.93) S = 4 (3.67) T = 3 (2.75)V = 5 (4.59) W = 2 (1.83)

Isoelectric Point (pI)	4.71
Net Charge at pH 7	-4.96
Charge	-4
Charge Attribute	Acidic
Antigenicity, Hydrophobicity & Hydrophilicity	
Secondary Structure Predicition	
Amphiphilicity Helix	E E E A E E A R G A G E A Q E R A V T A T A A A E E A A A G P G P V R L S A P R G C E S A A A E A A A A E E A A A G G G V G E P A P G R R G A E D E A A A W R SW R G G G R R A V A A A A A A A A P D P G G R A R T H V S R Legend: Helix
Chou-Fasman Prediction	E E E A E E A R G A G E A Q E R A V T A T A A A A E E A A A G P G P V R L S A P R G C E S A A A E A A A A E E A A A G G G V G E P A P G R R G A E D E A A A W R S WR G G G R R A V A A A A A A A A P D P G G R A R T H V S R Legend:Alpha,Beta, Turn
GOR-I Prediction	E E E A E E A R G A G E A Q E R A V T A T A A A E E A A A G P G P V R L S A P R G C E S A A A E A A A A E E A A A G G G V G E P A P G R R G A E D E A A A W R S W R G G G R R A V A A A A A A A A P D P G G R A R T H V S R Legend:Alpha,Beta, Turn, Coil

c) Characterisation of sPEP translated from *PTPN21* found in HEK cells

Information was obtained using a web-based peptide/protein analysis tool from LifeTein (http://lifetein.com/peptide-analysis-tool.html), indicating properties of the sPEP including protein hydrophobicity, secondary structure prediction, transmembrane region predictions, protein flexibility.

Primary information of sPEP translated from *PTPN21*	
Interpreted amino acid sequence	MGLYCHTGTITEYLSRFLADGMGTGNCNYSNGDSRR GGWKGEEL
Number of Standard Residues	44
Chemical Formula	$C_{204}H_{310}N_{60}O_{68}S_4$
Molecular Weight	4819.55 g/mol
Sequence Composition (in percentage)	Acidic: 11.36 Basic: 11.36 Neutral: 43.18 Hydrophobic: 34.09
Amino Acid Composition	A = 1 (2.27) C = 2 (4.55) D = 2 (4.55) E = 3 (6.82) F = 1 (2.27) G = 9 (20.45) H = 1 (2.27) I = 1 (2.27) K = 1 (2.27) L = 4 (9.09) M = 2 (4.55) N = 3 (6.82) R = 3 (6.82) S = 3 (6.82) T = 4 (9.09) W = 1 (2.27) Y = 3 (6.82)
Isoelectric Point (pI)	5.48

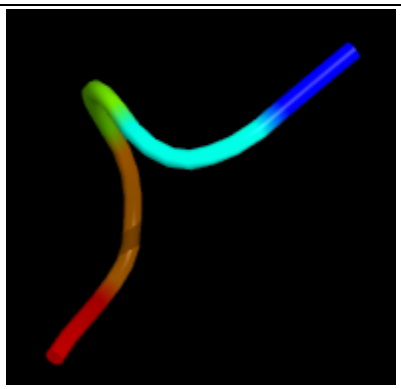
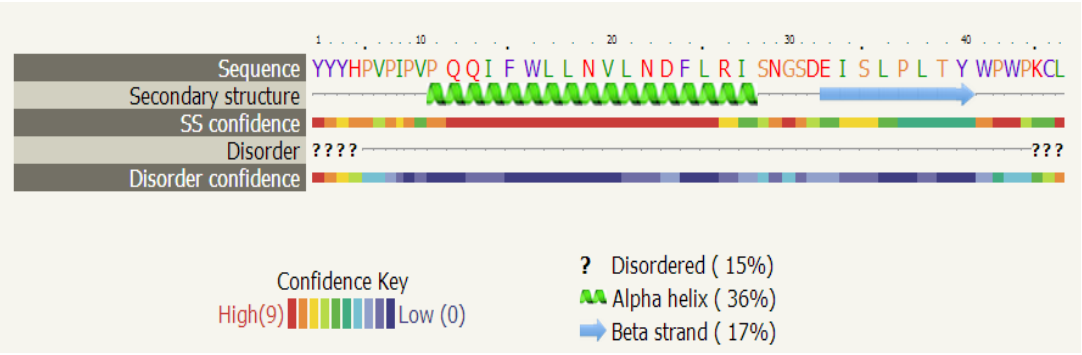
Net Charge at pH 7	-1.07
Charge	0
Charge Attribute	Neutral
Secondary Structure Predicition	
Amphiphilicity Helix	M G L Y C H T G T I T E Y L S R F L A D G M G T G N C N Y S **N G D S R R G G** W K G E E L Legend: **Helix**
Chou-Fasman Prediction	M G L Y C **H** T G T I T E Y **L S R F L A** D **G** M **G T G** N **C N Y S N G D S R R** G G **W KG E E L** Legend:**Alpha, Turn**
GOR-I Prediction	M G L Y C H T G **T I T E Y L S R F L** **A D G M G T G N C N Y S N G D S R R** G G W K G E E L Legend:**Alpha,Beta, Turn, Coil**

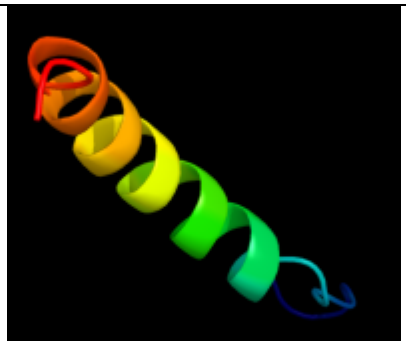
Appendix 5

a) Protein secondary structure prediction of the sPEPs found in HEK cells

Protein secondary structure prediction of sPEP translated from *LINC00950* non-coding RNA found in HEK cells

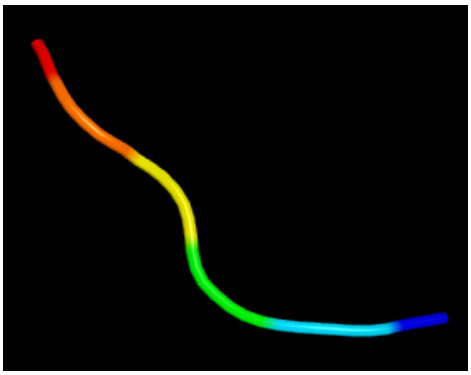
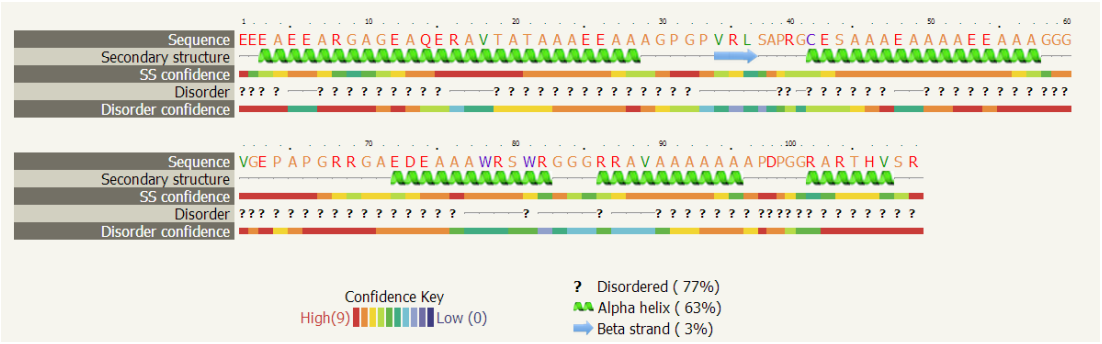
Information was analysed using Protein structure prediction on the web: a case study using the Phyre server (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index).

Protein structure prediction- sPEP translated from *LINC00950* non-coding RNA
      Confidence: 31.3% Coverage: 11% 5 residues (11% of the sequence) have been modelled with 31.3% confidence by the single highest scoring template.
Molecular description

	
Classification	Apoptosis
Structure Weight	3321.95
Molecule	Activator of apoptosis harakiri (Harakiri is a BH3-only member of the Bcl-2 family that localizes in membranes and induces cell death by bindin g toprosurvival Bcl-x(L) and Bcl-2.)
Polymer	1
Length	32
Chain	A

b) Protein secondary structure prediction of sPEP translated from *TM9SF3* found in HEK cells

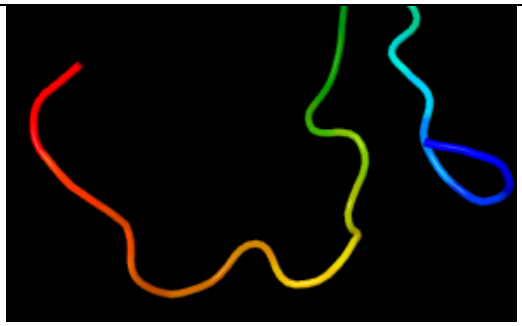
Information was analysed using Protein structure prediction on the web: a case study using the Phyre server (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index).

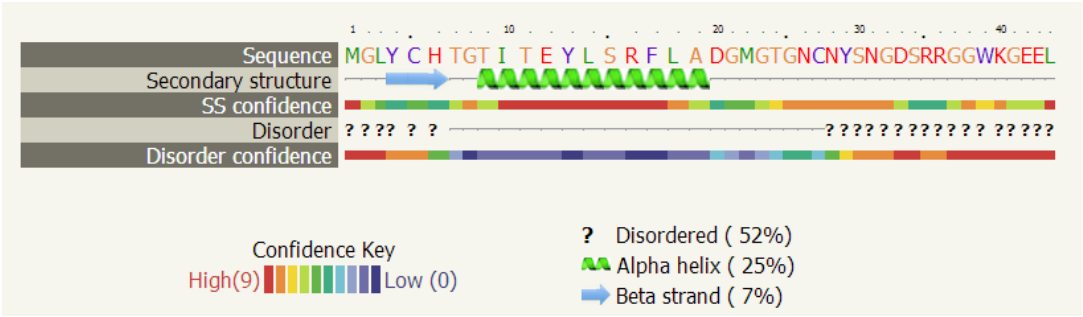
Protein structure prediction- sPEP translated from *TM9SF3*	
  Confidence: 5.1% Coverage: 6% 6 residues (6% of the sequence) have been modelled with 5.1% confidence by the single highest scoring template.	
PDB Molecule	putative uncharacterized protein
Chain	A

c) Protein secondary structure prediction of sPEP translated from *PTPN21* found in HEK cells

Information was analysed using Protein structure prediction on the web: a case study using the Phyre server (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index).

Protein structure prediction- sPEP translated from*PTPN21*





Confidence: 27.6% Coverage: 50%

22 residues (50% of the sequence) have been modelled with 27.6% confidence by the single highest scoring template.

Fold: 6-bladed beta-propeller

Superfamily: Sialidases

Family: Sialidases (neuraminidases)

Molecular description

PDB Description	structures of aromatic inhibitors of influenza virus neuraminidase
PDB Compounds	influenza a subtype n2 neuraminidase

Appendix 6

Protein localisaion prediction of the sPEPs found in HEK cells

a) Subcellular Localization Prediction for the sPEP translated from *LINC00950* non-coding RNA found in HEK cells

Information was analysed using Protein structure prediction on the web: The PredictProtein server Rost, LocTree2 (https://www.predictprotein.org/), predicting the subcellular localisation of all domains.

Protein Localisation Prediction- sPEP translated from *LINC00950* non-coding RNA	
Eukarya domain	Secreted
Prediction confidence	51

b) Subcellular Localization Prediction for the sPEP translated from *TM9SF3* found in HEK cells

Information was analysed using Protein structure prediction on the web: The PredictProtein server Rost, LocTree2 (https://www.predictprotein.org/), predicting the subcellular localisation of all domains.

Protein Localisation Prediction- sPEP translated from *TM9SF3*	
Eukarya domain	Cytoplasm
Prediction confidence	46

c) Subcellular Localization Prediction for the sPEP translated from *PTPN21* found in HEK cells

Information was analysed using Protein structure prediction on the web: The PredictProtein server Rost, LocTree2 (https://www.predictprotein.org/), predicting the subcellular localisation of all domains.

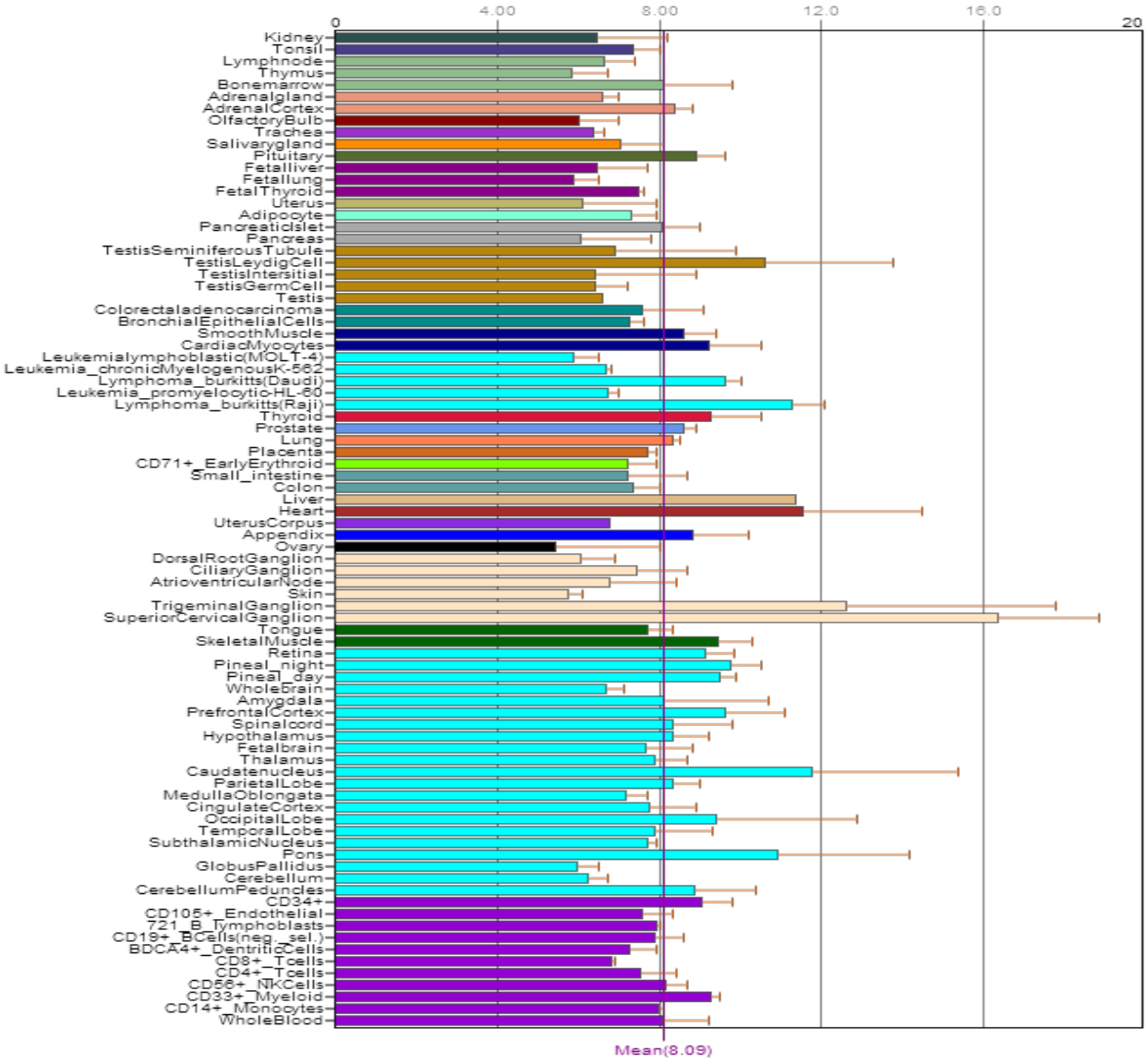
Protein Localisation Prediction- sPEP translated from*PTPN21*	
Eukarya domain	Secreted
Prediction confidence	44

Appendix 7

Gene and protein expression in a range of cell/tissue types of the sPEPs found in HEK cells

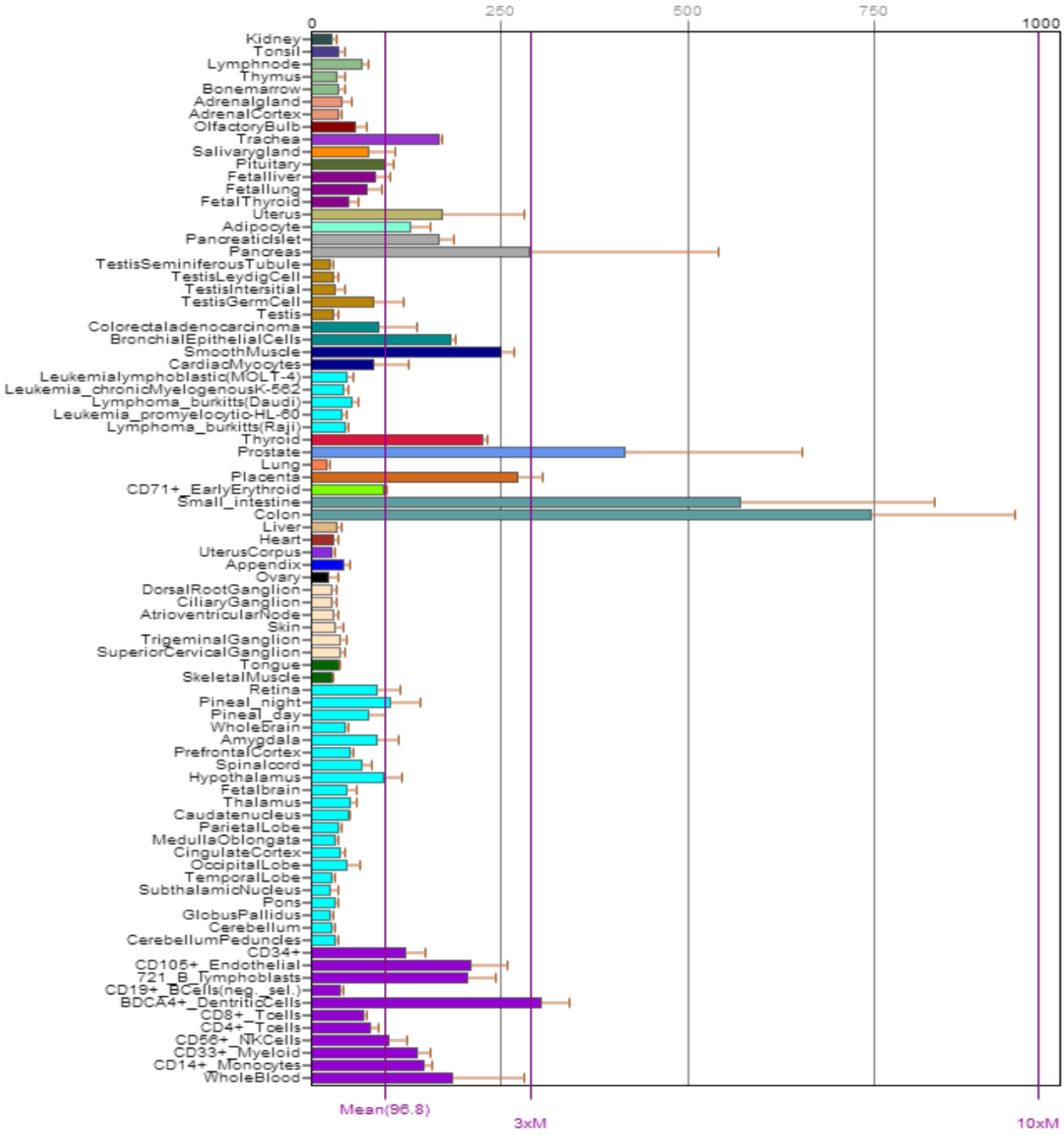
a) Gene and protein expression in a range of cell/tissue types of *LINC00950* non-coding RNA associated with the identified sPEP

Information were analysed using a web-based gene and protein function analysis tool from BioGPF (http://biogps.org). X-axis represents the expression level in relative units in different cell/tissue types shown in bars. Higher value indicates higher expression and gene activity. The highest expression is in the superior cervical ganglion (shown in light yellow bar).



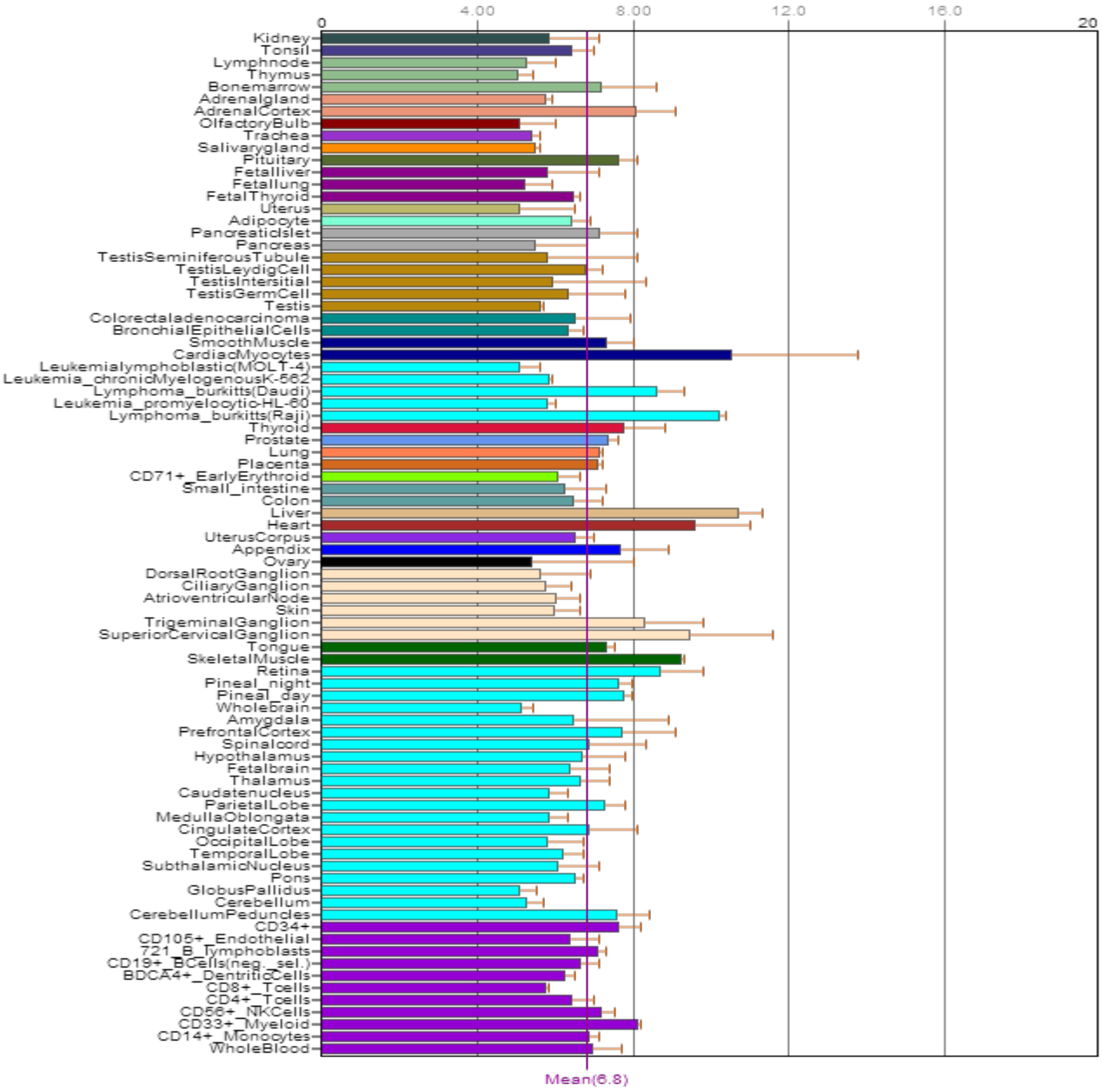
b) Gene and protein expression in a range of cell/tissue types of *TM9SF3* human mRNA associated with the identified sPEP

Information were analysed using a web-based gene and protein function analysis tool from BioGPF (http://biogps.org). X-axis represents the expression level in relative units in different cell/tissue types shown in bars. Higher value indicates higher expression and gene activity. The highest expression is in the colon (shown in blue green bar).



c) Gene and protein expression in a range of cell/tissue types of *PTPN21* human mRNA associated with the identified sPEP

Information were analysed using a web-based gene and protein function analysis tool from BioGPF (http://biogps.org). X-axis represents the expression level in relative units in different cell/tissue types shown in bars. Higher value indicates higher expression and gene activity. The higher expressions are in the cardiac myocytes (shown in dark blue bar), lymphoma (shown in light blue bar), and liver (shown in light brown bar).



Appendix 8

Motif prediction for the sPEPs found in HEK cells

a) Motif prediction for the sPEP translated from *LINC00950* non-coding RNA found in HEK cells

Information was analysed using MyHit, a free database website devoted to protein domains: http://myhits.isb-sib.ch/cgi-bin/motif_scan.

Motif Prediction Results- sPEP translated from*LINC00950* non-coding RNA		
Database of motifs	**Match score**	**Motif informationc**
PROSITE patterns (frequent match producers) [freq_pat]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:** 29-32	ASN_GLYCOSYLATION *N-glycosylation site*

b) Motif prediction for the sPEP translated from *TM9SF3* found in HEK cells

Information was analysed using MyHit, a free database website devoted to protein domains: http://myhits.isb-sib.ch/cgi-bin/motif_scan.

Motif Prediction Results- sPEP translated from*TM9SF3*		
Database of motifs	**Match score**	**Motif informationc**
PROSITE profiles	**Status:** A strong match: it is very unlikely that this match is a false	ALA_RICH *Alanine-rich region profile*

	positive **Position:** 9-46 **raw-score = 87** **N-score = 16.143** **E-value = 1.5e-09**	
PROSITE patterns (frequent match producers) [freq_pat]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:** 66-69; 84-87	AMIDATION *Amidation site*
PROSITE patterns (frequent match producers) [freq_pat]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:** 9-14; 41-46; 58-63; 84-89	MYRISTYL *N-myristoylation site*
PROSITE patterns (frequent match producers) [freq_pat]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:** 80-82	PKC_PHOSPHO_SITE *Protein kinase C phosphorylation site*

Pfam HMMs (local models) [pfam_fs]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:67-78** **raw-score = 2.1** **N-score = 8.285** **E-value = 0.11**	Gastrin *Gastrin/cholecystokinin family*

c) Motif prediction for the sPEP translated from *PTPN21* found in HEK cells

Information was analysed using MyHit, a free database website devoted to protein domains: http://myhits.isb-sib.ch/cgi-bin/motif_scan.

Motif Prediction Results- sPEP translated from*PTPN21*		
Database of motifs	**Match score**	**Motif informationc**
PROSITE patterns (frequent match producers) [freq_pat]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:** 28-31	ASN_GLYCOSYLATION *N-glycosylation site*
PROSITE patterns (frequent match	**Status:** Questionable or weak match: determining the true or false negative	CK2_PHOSPHO_SITE *Casein kinase II phosphorylation*

producers) [freq_pat]	status of this match requires additional biological evidences **Position:** 9-12; 30-33	*site*
PROSITE patterns (frequent match producers) [freq_pat]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:** 21-26; 37-42	MYRISTYL *N-myristoylation site*
PROSITE patterns (frequent match producers) [freq_pat]	**Status:** Questionable or weak match: determining the true or false negative status of this match requires additional biological evidences **Position:34-36**	PKC_PHOSPHO_SITE *Protein kinase C phosphorylation site.*

Appendix 9

Analysis of CGG-repeat uORFs in neural transcripts using uPEPperoni

To explore traces of conservation between human CGG-repeat uORFs in neural transcripts with different species, uORFs (a neural uORF list from John Carson's search group) containing CGG-repeat elements were analysed using uPEPperoni online search engine, which detects conserved uORFs in eukaryotic transcripts. 23 hits of conserved uPEPs in other species were obtained in total 226 human neural uORFs from the list.

Gene symbol	Gene name	CGG repeats	Location of CGG repeats	Chromosomal location of the CGG repeats	Conserved species
DEDD	Death effector domain containing	cgg	5'	Chr1: 159369054-159369074	Human, mouse, bos taurus, pongo abeli, macaca mulatt
B4GALT2	Betaglcnac beta 1,4-galactosyltransferase, polypeptide 2	cgg	5'	Chr1: 44217597-44217623	Human, mouse, bos taurus
RXRG	Retinoid x receptor, gamma	cgg	5'	Chr1: 163681015-163681035	Human, mouse, rat, bos taurus, pongo abeli, macaca mulatt
NRXN1	Neurexin 1	cgg	5'	Chr2: 50427515-50427541	Human, bos taurus,
TXNDC5	Thioredoxin domain containing 5	cgg	ORF	Chr6: 7855870-7855887	Human, danio rerio
HDGFL1	Hepatoma derived growth factor-like 1	cgg	ORF	Chr6: 22678328-22678345	Human, bos taurus,
PDGFA	Platelet-derived growth factor alpha polypeptide	cgg	5'	Chr7: 525595-525612	Human, mouse, canis rupus
ZNF212	Zinc finger protein 212	cgg	Promotor/ 5'	Chr7: 148567696-148567713	Human, ciona intestinalis
SLCO5A1	Solute carrier organic anion transporter family, member 5a1	cgg	5'	Chr8: 70907784-70907807	Human, mouse, rat

NTNG2	Netrin g2	cgg	5'	Chr9: 134027273-134027290	Human, rat, bos taurus, macaca mulatt
KCNMA1	Potassium large conductance calcium-activated channel, subfamily m, alpha member 1	cgg	ORF	Chr10: 79067351-79067371	Human, bos taurus,
FAM155A	family with sequence similarity 155, member A homolog 1 (drosophila)	cgg	ORF	Chr13: 107316518-107316535	Human, pongo abeli
ARHGEF7	Rho guanine nucleotide exchange factor (gef) 7	cgg	5'	Chr13: 110565785-110565802	Human, mouse, rat
PABPN1	Poly(a) binding protein, nuclear 1	cgg	5'	Chr14: 22860523-22860540	Human, bos taurus,
ARIH1	Ariadne ubiquitin-conjugating enzyme E2 binding protein homolog 1	cgg	ORF	Chr15: 70554271-70554291	Human, gallus gallus
SMG1	SMG1 phosphatidylinositol 3-kinase-related kinase	cgg	ORF	Chr16: 18844812-18844829	Human, xenopus laevis
ST3GAL2	ST3 beta-galactoside alpha-2,3-sialyltransferase 2	cgg	5'	Chr16: 69030429-69030452	Human, mouse, bos taurus, macaca mulatt
ZFHX3	Zinc finger homeobox 3	cgg	ORF	Chr16: 71379096-71379116	Human, mouse
ZBTB4	Zinc finger and btb domain containing 4	cgg	5'	Chr17: 7323646-7323663	Human, mouse
WIPF2	WAS/WASL interacting protein family, member 2	cgg	5'	Chr17: 35629149-35629175	Human, sus scrofa
TCF4	Transcription factor 4	cgg	5'	Chr18: 51406386-51406406	Human, mouse
CAPNS1	Calpain, small subunit 1	cgg	ORF	Chr19: 41323783-41323800	Human, oryctolagus cuniculus
DSCAM	Down syndrome cell adhesion molecule	cgg	5'	Chr21: 41140576-41140593	Human, xenopus laevis